GODFREY, KELLY ELIZABETH, Ph.D. A Comparison of Kernel Equating and IRT True Score Equating Methods. (2007)
Directed by Dr. Terry A. Ackerman. 181 pp.

This two-part study investigates 1) the impact of loglinear model selection in pre-smoothing observed score distributions on the kernel method of test equating and 2) the differences between kernel equating, chained equipercentile equating, and true score methods of concurrent calibration and Stocking and Lord's transformation method. Data were simulated to emulate realistic situations in which test difficulty differed, sample sizes varied, anchor test lengths were of varying lengths, and test lengths ranged from 20 items to 100 items. Difficulty of anchor tests were held constant. Because data were simulated in a single group (SG) format, traditional unsmoothed equipercentile equating was used as a criterion by which all other methods, which use the non-equivalent groups with an anchor test design (NEAT), were compared. Data were simulated using IceDog (ETS, 2007) and analyzed using KE software (ETS, 2007), MULTILOG (Thissen, 2003), IceDog (ETS, 2007), PARSCALE (Muraki & Bock, 2003) and Fortran programming code developed by the author. Results indicate the impact of equating technique chosen on examinees' test scores in a variety of realistic situations, and have further recommendations for further study.

A COMPARISON OF KERNEL EQUATING AND IRT TRUE SCORE EQUATING METHODS

by

Kelly Elizabeth Godfrey

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2007

Approved by

_____
Committee Chair

For my family.

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School a The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I'd like to thank my committee, first and foremost, for their guidance, patience, and incredible encouragement during the past year. Dr. Ackerman, you have taught me more than I ever dreamed I could learn, and I thank you so much. Alina, thank you for your time and patience, and for taking me under your wing to help me learn an immense amount of material in a relatively short time period. To Rick Morgan, I enjoyed your class more than I ever let on, and your advice and lessons "from the trenches" were priceless. To Bob Henson, you've done a great job as a professor, and I hope to someday be as successful at making difficult tasks look so easy. I want to thank Henry Chen and Fred Robin at ETS, who both provided invaluable technical support and answered many questions. I also have to express my appreciation for Mike Nering, Liz Burton, and the gang at Measured Progress, who first introduced me to equating and strengthened the passion for psychometrics and test quality. Thank you all for your patience and dedication!

To my friends, Lacy, Anne, Wenmin, Yingchen, Nichole, Lynnette, and Athena: I couldn't have gotten through graduate school without your support, phone calls, visits, Anne's random packages, and Wenmin's incredible cooking! Whenever I needed to vent frustrations, ask questions, request advice, or just "talk shop", you all were there listening, or at least trying to feign interest while I talked through technical problems surrounding this dissertation. You are the greatest!

And last but not least, to my father: I never would have had the courage to follow my dreams if I didn't have you behind me 100 percent. You taught me to hold my head up and to keep going, even when times got tough and giving up looked so easy. Your support and undying dedication to my education will never go unappreciated. You are a true hero, and I wouldn't be here today without you. I love you.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The American Educational Research Association standards (1999) call for common score scales over time so that periodic investigations of scale stability can be conducted (standard 4.17). The standards also state: "a clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably" (standard 4.10). In order to maintain security, many tests have multiple forms assigned randomly to examinee groups across multiple administration dates, times, and locations. Therefore, according to the standards for test development, actions must be taken to scale scores so that they are considered interchangeable.

Equating is essential for most, if not all, testing programs. Because there is no way to guarantee that the test forms are the same difficulty, testing programs must equate them to ensure that the scores are comparable, meaning that it does not matter which test form the examinee completed. There are numerous approaches to test equating; two major types being observed score equating and true score equating. Observed score methods include, but are not limited to, equipercentile equating, chained equipercentile equating (Kolen & Brennan, 1995), and the latest approach, kernel equating (von Davier, Holland, & Thayer, 2004). True score equating methods include concurrent calibration

(Lord, 1980), and Stocking and Lord's (1983) transformation method, also referred to as TBLT (transforming B's using a least squares technique).

IRT True Score Equating Techniques

Concurrent calibration is conducted by estimating the item parameters for each form simultaneously using an estimation program such as MULTILOG (Thissen, 2003). The item parameters are estimated using all of the examinees, ignoring missing responses for those items an examinee did not receive (Lord, 1980). An advantage of using this method for item parameter estimation is that the estimates are automatically on the same scale. However, it has its disadvantages. For instance, when the ability distributions differ dramatically, the specification of the population parameters becomes difficult (Kim & Cohen, 1998).

The Stocking and Lord approach (1983) is classified as a characteristic curve method. It considers all estimated parameters simultaneously, focusing on differences between test characteristic curves (Kolen & Brennan, 2004). This method involves item parameter estimates that have been calibrated separately, and transforms the scale of the parameters for Form *X* onto the scale of Form *Y* using the common (or anchor) items. This method overcomes the obstacle of having two items with very different b-parameter estimations producing the same item characteristic curve, but does not take into account error in the parameter estimations. When sample sizes are small, or differ drastically in size, this may present a problem in equating (Kolen & Brennan, 2004).

Other popular true score approaches include the mean/sigma method (Kolen & Brennan, 2004), and the robust mean/sigma method (Linn, Levine, Hastings & Wardrop, 1981) and the fixed common item parameter method. The mean/sigma and robust mean/sigma approaches use the means and standard deviations (sigma) of the difficulty parameters in the anchor tests to transform scales in the common-item design. The fixed common item parameter method, or FCIP, estimates the parameters of the items in the anchor test along with the items in form Y, then fixes those parameter estimates when calibrating form *X* and the anchor. Although these methods are widely used, they are not investigated here.

Observed Score Equating Techniques

Observed score equating is a common practice among major testing programs. Raw total scores are calculated for each examinee and are used to create a score distribution for each test form. These methods are simpler than the true score methods mentioned above, and include both linear and equipercentile techniques. Observed score equating is devised of two components: the data collection design (also called the design function) and the equating method (von Davier, Holland, & Thayer, 2004). The studies presented here focus on two design functions: single group (SG) and non-equivalent groups with an anchor test (NEAT) and equipercentile methods.

One common observed score method is equipercentile equating. This technique uses percentile rankings to scale scores from test form *X* to the scale of test form Y. A score of *x* on form *X* equates to a score of *y* on form *Y* if they have the same percentile

rank (Petersen, Cook, & Stocking, 1983). An advantage of equipercentile equating is that it does not assume that the difference in difficulty between two test forms is uniform across the score scale (Kolen & Brennan, 2004).

A variation of the traditional equipercentile equating is chained equipercentile equating (Angoff, 1971), which is used in the non-equivalent groups with an anchor test (NEAT) design. This method uses the common items in the anchor test to equate the unique items on form *X* for population *P* to the unique items on form *Y* for population *Q* by a chain of conversions (Kolen & Brennan, 2004). Despite its simplicity, this method has its advantages; it is considered accurate and stable in practice (Livingston, Dorans, & Wright, 1990). A related observed score equating technique is the chained linear method, which is simply a special case of the equipercentile approach (von Davier et al., 2004).

Other popular observed score methods include Tucker (Gulliksen, 1950) and Levine methods (Kolen & Brennan, 2004). Both of these are linear methods, which condition on the anchor test (used in the NEAT design). The Tucker method creates a synthetic distribution of scores for group *P* on test form *Y* and group *Q* on form *X* and uses regression to calculate the relationships between unique test forms and the anchor (Kolen & Brennan, 2004). However, if the relationship between the anchor and the unique forms *X* and *Y* are not strong, then this equating method delivers fallible results. This method has an analogous equating method within kernel equating: post-stratification equating, or PSE (von Davier et al., 2004). The Levine method, on the other hand, relates observed scores on the two forms to each other, but its assumptions focus on true scores,

rather than observed and can be a strong equating technique when group differences are large (Kolen & Brennan, 2004).

Introduction of the Kernel Method

Kernel equating, a method of test equating developed by Paul Holland and Dorothy Thayer (1987), and later updated along with Alina von Davier (2004), utilizes a five-step process for equipercentile equating where linear equating is a special case. Output includes the standard error of equating (SEE), which measures the equating function sampling variability, the standard error of equating difference (SEED), which measures the standard error of the difference between two equating functions, as well as the equated results and the comparisons of moments, cumulative distribution functions, and the bandwidths.

Kernel equating uses a five step process: (1) pre-smoothing, which uses a loglinear model to smooth the distributions of observed scores; (2) estimation of score probabilities, which uses the design function to calculate the score probabilities **r** (for test form *X*) and **s** (for test form *Y*); (3) continuization, a step similar to Kolen and Brennan's "post-smoothing" (2004, von Davier, et al., 2004), in which continuous cdf's are calculated from the discrete cdf's, and bandwidth parameters are chosen; (4) equating, where the continuous cdf's from step 3 are used to automatically create an equipercentile equating function; and finally, (5) calculation of the standard error of equating, as well as the standard error of the equating difference. These steps are explained further in Chapter 2.

Advantages of kernel equating over other methods include the thorough diagnostics provided in the output, allowing researchers to judge whether or not the equating function is naturally linear (as opposed to forcing a linear relationship in the choice of equating technique); the more realistic assumptions as compared to other methods, such as IRT-based methods, and relaxed restrictions on the data.

Research Purposes and Questions

The purpose of the studies presented here is twofold: to investigate the impact of the loglinear model chosen to pre-smooth the score distributions, and to compare kernel equating methods to the more traditional methods of chained equipercentile equating and the IRT methods using the NEAT design, concurrent calibration and Stocking and Lord to each other and to a criterion equating.

Pre-smoothing is an important step in the kernel equating process, yet the impact of the loglinear model chosen in this step has yet to be investigated. The first study presented here involves seven different models used to pre-smooth the score distributions: 4-4-0, 4-4-1, 4-4-4, 6-6-4, 8-8-4, 10-10-4, and 12-12-4. The three numbers representing each model are the number of moments preserved for test form *X*, the number of preserved moments for *Y*, and the number of cross-product moments between the two forms, respectively. These models are explained further in Chapter 2. An important property of estimating score probabilities is preserving the integrity of the raw score distribution when possible (von Davier, et al., 2004). A larger model (12-12-4, for instance) reflects the raw score distribution more closely. However, the approach to pre-

smoothing should be to fit the data as well as possible with as few parameters as possible (von Davier, et al., 2004). It is a goal of this study to investigate the effect on equating results of the loglinear model fit to the raw data.

The second study presented here is a comparison of two observed score equating methods, kernel and traditional chained equipercentile, and two IRT methods, concurrent calibration and Stocking and Lord. With varying factors such as test form length, anchor length, examinee group ability differences, and sample sizes, the four equating methods are compared using a NEAT design to each other and to a criterion equating. Data are simulated using a two-parameter logistic (2-PL) model in ICEDOG (ETS, 2006). Details of the simulation are presented below in Chapter 3. This study is among the first, to date, to apply kernel methods of equating to simulated data and compare the resulting functions to more traditional and established methods. It is a goal of this study to further the understanding of the usefulness and appropriateness of the kernel equating method, as well as the differences and similarities to and among other equating techniques.

CHAPTER II

REVIEW OF THE LITERATURE

This chapter presents literature related to the observed score equating techniques traditional chained equipercentile and kernel equating and the true score methods concurrent calibration and Stocking and Lord. It first lists general guidelines for equating, then explains design functions and the role they play on equating. This chapter also explains the four equating techniques of interest here and how they are calculated. Finally, relevant studies investigating and comparing the equating techniques are discussed, and arguments are made for the additions this research can add to the knowledge about kernel equating.

Equating Guidelines

Equating methods were developed to account for differences in difficulty of items on test forms measuring the same construct(s) (Kolen & Brennan, 2004; Cook & Eignor, 1991). However, these approaches, including true score methods, were not designed to account for large differences in difficulty, differences in content, or large differences in reliability (Cook et al., 1991). Several researchers have illustrated requirements, or guidelines, for test equating (Lord, 1980; Angoff, 1971; Kolen & Brennan, 2004). Angoff (1971) described four requirements in order for two test forms to have been successfully equated: (a) the same ability needs to be measured in each of the test forms,

(b) the scale score conversion should be independent of the data used and should be generalizable to other similar sets of examinees, (c) equated scores should be interchangeable, and (d) the equating must be symmetric. Likewise, Lord (1980) listed three requirements true score equating unidimensional tests of the same construct: (a) equity, (b) group invariance, and (c) symmetry. Kolen and Brennan (2004) list similar requirements. Von Davier, Holland, and Thayer (2004) list Dorans and Holland's (2000) explanation of five requirements for equating, and those will be described here.

The first requirement is that of equal constructs. This means that the test developer oftentimes works with a test blueprint to assure that each of the forms meet set specifications and contain the same format of items measuring the same construct. This includes making the tests the same length, approximately the same difficulty, and designed for the same audience.

The second requirement listed by Dorans and Holland (2000) is that of equal reliabilities. Lord (1980) cites this requirement in detail as Theorem 13.3.1 (p. 198): "Under realistic regularity conditions, scores *X* and *Y* on two tests cannot be equated unless either (1) both scores are perfectly reliable or (2) the two tests are strictly parallel." Although the test forms studied here do not have a specified construct, as they are simulated, they are unidimensional in nature and are designed to be equally reliable.

The third requirement listed by Dorans and Holland (2000) is that of symmetry; the equating function for *X* to *Y* must be the inverse of the function for *Y* to X. To explain further, if a score on form *X* equates to a particular score on form Y, then the score on *Y* should equate back to the original score on X. For example, if a 200 on form

X equates to a 220 on form Y, then a score of 220 on *Y* should reverse-equate to a 200 on X. It is important to note that equating is not the same as prediction, and statistics such as regression are not necessarily symmetric in nature.

The fourth requirement from Dorans and Holland (2000) is equity. It should not matter to the examinee which test form he or she completed. Lord (1980) also lists this requirement, explaining that the conditional frequency distributions of the equated scores of *X* to *Y* must be equal to those of *X* for each ability level. Equating accounts for varying difficulties between multiple test forms of the same construct, so this requirement is an integral part of a successful equating.

Dorans' and Holland's (2000) final requirement is population invariance. This means that the equating function, $x(y)$ should remain the same regardless of the population used to determine it. If this requirement holds, the equating results should be generalizable to other similar populations for which the tests have been designed.

Design Functions

Equating requires some sort of commonality, whether it between test items or test takers. This is specified in the design function, which describes how the data were collected. The choice of equating approach is influenced by the design function. Some commonality must exist between the two test forms and examinee groups, whether that is in common items or common examinees. There are four basic designs: single group (SG), counterbalanced (CB), equivalent groups (EG) and non-equivalent groups with an anchor test (NEAT). For the purposes of the research presented here, this chapter will

focus mostly on two design functions: SG and NEAT, but for the sake of comprehensiveness, the other two, CB and EG, are discussed briefly.

In the SG design, all examinees take both test forms, *X* and Y. The order of test form administration remains uniform for all examinees (alternating the order the forms are administered for the group is termed the counterbalanced, or CB, design). This design, although simple, can be expensive to implement and unrealistic in the assumption that order effects and test fatigue do not have an effect on test scores, and is not recommended in practice (Kolen & Brennan, 2004). Tables 2.1 shows the SG design: one examinee group takes both test forms, *X* and Y.

Table 2.1: Single Group Design

	X	*Y*
Examinee Group	✓	✓

The counterbalanced, or CB, design is greatly similar to the SG design. However, whereas the SG design does not take order effects into account in the administration of test forms, the CB does. Half of the examinee group is given test form *X* first, and then test form *Y* second. The other half of the examinee group takes the same two tests in reverse order. Table 2.2 below demonstrates this. This design is essentially treated as the combination of two SG designs, because the same examinees take both forms, *X* and *Y* (Kolen & Brennan, 2004; von Davier et al., 2004).

Table 2.2: Counterbalanced Design

	X_1	Y_2	Y_1	X_2
Examinee Group *P*	✓	✓		
Examinee Group *Q*			✓	✓

In the NEAT design, two examinee groups, *P* and *Q*, take test forms *X* and Y. Because the groups are not assumed equivalent, and they are not taking the same test forms, they must be linked through an anchor test, or test that is similar to the test forms to be equated that is used to equate the test forms and account for group differences in ability. The target population, *T*, is the combination of *P* and *Q*, and is defined as $T = wP + (1 - w)Q$. If *P* and *Q* are equal in size, *w* is equal to 0.5. Table 2.3 below shows the NEAT design, demonstrating the relationship between examinee group and test form. *P* and *Q* are not assumed equivalent in any way.

Table 2.3: Non-Equivalent Groups with an Anchor Test Design

		X	*Y*	*A*
T	Population *P*	✓		✓
	Population *Q*		✓	✓

The final equating data collection design is the equivalent groups, or EG, design. Here, two separate groups of examinees, *P* and *Q*, are administered test forms *X* and *Y*, respectively, and are assumed to be equivalent groups in ability of the test construct. This design can be obtained by the random assignment of test forms, such as spiraling, so that the examinee groups are randomly equivalent.

There are two major approaches of equating: observed-score and true-score. True score equating methods include concurrent item response theory (IRT) calibration (Lord, 1980) and Stocking and Lord's transformation method, or TBLT (transforming B's using the least squares technique) (Stocking & Lord, 1983). Observed score equating includes linear techniques such as chained linear equating, equipercentile methods such as chained equipercentile, and kernel equating. All of these methods are discussed below.

True Score Equating

Concurrent Calibration

Concurrent calibration involves the simultaneous calibration of all items on both test forms using both examinee groups (Lord, 1980; Kolen & Brennan, 2004). The two-parameter logistic (2-PL) model will be used in this study. In this model, two item parameters, discrimination (*a*) and difficulty, or location, (*b*) are estimated, as well as one latent examinee ability parameter, θ.

The *a*-parameter indicates how well an item discriminates between those who have at least the ability level or higher than required to answer the item and those who do not. It is the slope of the item characteristic curve (ICC), which is a generally non-linear curve that plots performance on an item against ability, or θ (Hambleton & Swaminathan, 1985). A higher *a*- value indicates a steeper slope, and, thus, higher discrimination.

The *b*-parameter describes the item's difficulty level. *B*-parameters typically range from -2.0 to +2.0 (Hambleton & Swaminathan, 1985), but can certainly fall outside of that range, with lower values (below zero) indicating an easier item. The *b*-parameter

is also called the location parameter, and denotes the θ-value corresponding to the point of inflection in the ICC, where 50 percent of the examinees are likely to answer the item correctly.

The item characteristic curve for the 2-PL model is calculated as:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

where D is the scaling factor, 1.702 (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991).

The examinee parameter, θ, represents a subject's latent ability. Ability estimates are on the same metric as the *b*-parameters, and typically fall between -3.0 and +3.0.

In concurrent calibration in the NEAT design, the item parameters for all items are estimated using data from both examinee groups, ignoring missing data where an examinee did not see an item. Software such as MULTILOG (Thissen, 2003) calibrates the parameter estimates using an estimation technique called marginal maximum likelihood estimation (MMLE). This method calculates the likelihood function as:

$$L(u_1, u_2, ..., u_N \mid \theta, a, b) = \prod_{i=1}^{N} \prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$$

where N is the number of examinees, n is the number of items, u_i is the response vector for examinee i, P is the probability of getting the item correct, and Q is the probability of getting the item incorrect, or, $1 - P$ (Hambleton et al., 1991).

Stocking and Lord Transformation Method

In 1983, Stocking and Lord proposed a new method of transforming parameters to the same scale. Classified as a "characteristic curve method", their approach takes both item difficulty and discrimination into account by estimating item parameters for each test form separately, then calculating the test characteristic curves. Group *P* is calibrated for items on form *X* and the anchor items, while group *Q* is calibrated for items on form *Y* and the anchor items. Test characteristic curves, or the sum of the probability associated with each item's item characteristic curve (ICC), are calculated for each form, and then used to create an equating function from form *X* to form Y.

Observed Score Equating

Linear equating uses standard deviations to convert scores from one form to the other. Specifically, raw total scores that are the same distance from the mean in standard deviation units are set to be equal. This allows for test forms to vary in both difficulty and scale (Kolen & Brennan, 1995). Equipercentile equating uses cumulative distributions of raw total scores to equate two test forms, *X* and *Y* (Kolen & Brennan, 1995). In special cases where the two test forms have the same distributions, equipercentile equating and linear equating will be the same. Kernel equating is considered to be a special case of the equipercentile equating method (von Davier, et al., 2004).

Equipercentile Equating

In the SG design, the equipercentile equating function, as specified by Braun and Holland (1982) and summarized by Kolen and Brennan (2004), is the following:

$$e_Y(x) = G^{-1}[F(x)],$$

where G^{-1} is the inverse of the cumulative distribution function G, which is the cumulative distribution function of Y in the population. Likewise, because of the symmetry requirement in equating,

$$e_X(y) = F^{-1}[G(y)].$$

To calculate an equipercentile equating, one must use percentile ranks, that are the percentages falling below a certain specified score, then defining $F(x)$ as the "proportion of examinees in the population earning a score *at or below* x" (Kolen & Brennan, 2004, p. 44). We then calculate the inverse of the percentile rank:

$$P^{-1}[P^*] = \frac{P^*/100 - F(x_U^* - 1)}{F(x_U^*) - F(x_U^* - 1)} + (x_U^* - .5)$$

This expression is used to approximate (i.e., continuize) $F(x)$ and $G(y)$ from discrete score distributions to cdf's, so that the equating function may be calculated.

Chained Equipercentile Equating

Originally developed by Angoff in 1971 and later termed "chained equipercentile equating" by Livingston, Dorans and Wright in 1990 (Kolen & Brennan, 2004), this method converts scores collected using a NEAT design on Form X to scores on Form A (the common items referred to as the anchor), and then converts Form A scores to scores

on Form *Y* (Kolen & Brennan, 1995) using equipercentile methods. This method is simply a series of equipercentile equatings (like in the SG design) in a chained manner, using the anchor test to link the unique items of form *X* to the items in form Y.

Although this method is computationally simpler than some of its counterpoints, it is not without faults. One major concern regarding chained equipercentile equating is in regards to the equating of a long test (form X, Y) to a generally shorter test, A. Tests that are of different lengths are most likely not equally reliable, which violates a general guideline for equating. Also, the scores cannot be used interchangeably, which violates another of the requirements (Kolen & Brennan, 2004). However, this method does not require that the populations be very similar in ability, which makes this an attractive option for test equating.

Chained equipercentile equating has been touted as a stable and accurate method of observed score equating (Kolen & Brennan, 2004; Livingston et al., 1990). In 1993, Livingston suggested that the stability of this method may be improved by incorporating a loglinear smoothing of the joint distributions of scores.

Kernel Equating

Kernel equating, a method of test equating originally conceived by Holland and Thayer (1981) and further developed with von Davier (2004) , utilizes a five-step process (pre-smoothing, estimation, continuization, equating, calculating standard error of equating) for equipercentile equating where linear equating is a special case. Output includes the standard error of equating (SEE), the standard error of the equating

difference (SEED), as well as the equated results and the comparisons of moments, cumulative distribution functions, and the bandwidths.

Step 1: Pre-smoothing

The first step for kernel equating is pre-smoothing, using a loglinear model, the discrete raw score distribution. In both the SG and NEAT designs, these are bivariate distributions. According to von Davier, Holland, and Thayer (2004), there are four properties of concern when choosing an appropriate model for pre-smoothing: (1) consistency, that implies that as the sample size increases, the estimates should approach the population parameters; (2) efficiency, which means that the sample data should be as close to the population data as possible; (3) positivity, which means that every possible score should have a positive non-zero score probability, except in the case of the NEAT design with an internal anchor, where there are score combinations that are impossible; and (4) integrity, which means that the smoothed score distribution should match the observed score distribution in as many moments as possible. Von Davier, Holland, and Thayer (2004) recommend, based on their experience, that at least the first five or six moments of the test forms' raw score distributions be preserved.

The steps for smoothing a univariate distribution are different from those taken to smooth a bivariate distribution. Because both the SG and NEAT designs have bivariate data, that process will be the focus here. Information on the univariate loglinear model can be found in von Davier, Holland, and Thayer (2004) and Holland and Thayer (2000).

The bivariate loglinear model is expressed as:

$$\sum_{i,j} x_i^a y_j^b (f_{ij} / N),$$

where f_{ij} is the frequency of examinees with a score of x_i on one form and a score of y_j on the other. When *a* and *b* are both non-zero, the cross-product moments are being maintained in the model as well (Holland & Thayer, 2000). If no cross-product moments are preserved, then the relationship between the two test forms is ignored.

To choose the appropriate model, there are several fit statistics involved, the object being to establish the most appropriate fit with the fewest number of parameters. In the bivariate case, the best model for each univariate distribution is chosen, then the cross-product moments are added into the model and a bivariate loglinear model created. The likelihood ratio chi-square statistic gives an overall estimate of model fit, along with the Pearson chi-square and Freeman-Tukey chi-square statistics. These statistics do not always agree, as the Pearson chi-square can be larger when there are a lot of score combination possibilities with low frequencies (Holland & Thayer, 2000).

Another major fit statistic is the Freeman-Tukey residuals, which should show no pattern across the score scale, and should be relatively small, between -2.0 and +2.0 (von Davier et al., 2004). However, these are not very useful in showing model fit unless there are a lot of data where there are few zero frequencies. In most real cases, Holland and Thayer (2000) recommend that it is sufficient to compare the first three moments, location, scale, and skew, between the observed and fitted conditional distributions. Freeman-Tukey residuals give an overall estimate of the consistency and stability of the smoothed model (von Davier et al., 2004).

Other fit statistics obtained when pre-smoothing in kernel equating include the Akaike Information Criterion (AIC), which is calculated as:

$$AIC = 2k + G^2,$$

where k is the number of parameters estimated by the model, and G^2 is the likelihood ratio chi-square; as well as the related Bayesian Information Criterion (BIC); and Consistent Akaike Information Criterion (CAIC).

In addition to the loglinear model, this step also produces a covariance matrix that is used for computing the standard error of equating (SEE) and the standard error of the equating difference (SEED) described below in step 5. These "**C**-matrices" are calculated as:

$$\Sigma_{v(\hat{P})} = C_P C_P^t$$

and

$$\Sigma_{v(\hat{Q})} = C_Q C_Q^t$$

Because it is assumed that the samples are selected randomly and are independent of each other, the covariance between these two vectors of probability estimates is assumed to be zero. This leads to a joint **C** matrix of:

$$\Sigma_{v(\hat{P}),v(\hat{Q})} = \begin{pmatrix} C_P & 0 \\ 0 & C_Q \end{pmatrix} \begin{pmatrix} C_P & 0 \\ 0 & C_Q \end{pmatrix}^t$$

C is a $J \, X \, J$ matrix, where J indicates the number of possible scores on a form.

Step 2: Estimation of score probabilities

This step involves estimating the score probabilities from the distributions from pre-smoothing, according to the design function. Here, the SG and NEAT chained equating processes will be discussed.

The score probabilities for *X* and *Y* are denoted as ***r*** and ***s***. The design function "maps the population scores probabilities relevant to the data collected in a design into ***r*** and ***s***" (von Davier et al., 2004, p. 53). For the SG design, this is:

$$\begin{pmatrix} r \\ s \end{pmatrix} = DF(P) = \begin{pmatrix} M \\ N \end{pmatrix} v(P),$$

where *P* is the bivariate score probability matrix, and M and N are matrices of zeros and ones used to obtain *P* from its vector version, *v*(**P**). In the NEAT design with chain equating, however, this same step is calculated as two SG designs, such that:

$$\begin{pmatrix} r_P \\ t_P \end{pmatrix} = DF_P(P) = \begin{pmatrix} M_P \\ N_P \end{pmatrix} v(P)$$

and

$$\begin{pmatrix} t_Q \\ s_Q \end{pmatrix} = DF_Q(Q) = \begin{pmatrix} N_Q \\ M_Q \end{pmatrix} v(Q).$$

Here, *P* is the bivariate score probability matrix for population *P*, and *Q* is the bivariate score probability matrix for population *Q*. These are combined to give:

$$\begin{pmatrix} r_P \\ t_P \\ t_Q \\ s_Q \end{pmatrix} = DF(P,Q) = \begin{pmatrix} DF_P(P) \\ DF_Q(Q) \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} M_P \\ N_P \end{pmatrix} & \begin{matrix} 0 \\ \end{matrix} \\ \begin{matrix} 0 \\ \end{matrix} & \begin{pmatrix} N_Q \\ M_Q \end{pmatrix} \end{pmatrix} \begin{pmatrix} v(P) \\ v(Q) \end{pmatrix}.$$

This is calculated in this manner because chain equating links form X to anchor A, and then links anchor A to form Y (von Davier et al., 2004).

Step 3: Continuization

This third step in the kernel equating process involves Gaussian kernel smoothing, the term from which kernel equating gets its name. The data distributions at this point are still stepwise, as the data are still discrete. In order to compute inverse functions, as well as the basic equating function, the stepwise distribution needs to be smoothed into a continuous cumulative distribution function. In traditional equipercentile equating, linear interpolation is used with the midpoints of each score interval. However, in kernel equating, this step first involves computing the cdf's $F(x)$ and $G(y)$ as:

$$F(x) = \text{Prob}(X \leq x) = \sum_{j, x_j \leq x} r_j$$

and

$$G(y) = \text{Prob}(Y \leq y) = \sum_{k, y_k \leq y} s_k \ .$$

This step in kernel equating involves calculating bandwidths, h_X and h_Y. These positive numbers are used to Gaussian kernel smooth the cdf, such as:

$$F_{h_X}(x) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X} \right)$$

where

$$a_X = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + h_X^2}}.$$

The larger the bandwidth (h_X is greater than $10\sigma_X$), the closer to normal the continuized distribution of scores becomes, such that

$$F_{h_X}(x) \approx \Phi\left(\frac{x - \mu_X}{\sigma_X}\right)$$

(von Davier et al., 2004). When bandwidths are large, this leads to linear equating. When they are small (e.g., $h_X = 0.33$), the result is a density function similar to the original distribution: discontinuous with jumps. Kernel equating offers the option of automatically selecting the bandwidths by minimizing

$$PEN_1(h_X) = \sum_j (\hat{r}_j - \hat{f}_{h_X}(x_j))^2$$

and

$$PEN_2(h_X) = \sum_j A_j(1 - B_j)$$

where there is a penalty of 1 every time the density function is "U-shaped" around the score point. The two penalties are combined to give:

$$PEN_1(h_X) + K \times PEN_2(h_X)$$

where the larger K is, the fewer modes the distribution contains (von Davier et al., 2004). The resulting bandwidths are "optimal" and lead to more smooth Normal density functions. This is particularly useful when there are gaps, or "teeth" in the score distributions. Von Davier, Holland and Thayer (2004) recommend that in order for the

standard errors to be valid, the model has to fit, but there is no valid reason to keep these gaps, and the penalty function in this step removes them.

Step 4: Equating

This step is relatively straightforward. Using kernel equating's definition of equipercentile equating,

$$\hat{e}_Y(x) = e_Y(x;\hat{r},\hat{s}) = G_{h_Y}^{-1}(F_{h_X}(x;\hat{r});\hat{s}) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x)).$$

Because the density functions $\hat{F}_{h_X}$ and $\hat{G}_{h_Y}$ have been made continuous, calculating the inverses $\hat{F}_{h_X}^{-1}$ and $\hat{G}_{h_Y}^{-1}$ is easier (von Davier et al., 2004). Once the scores are equated, the moments are compared by calculating the percent relative error of the *p*th moment, or PRE(*p*). This can be found by calculating the moments of $\boldsymbol{Y}$ and $e_Y(\boldsymbol{X})$ as:

$$\mu_P(Y) = \sum_k (y_k)^p s_k$$

and

$$\mu_P(e_Y(X)) = \sum_j e_Y(x_j))^p r_j .$$

Then the PRE(*p*) is calculated as:

$$PRE(p) = 100\frac{\mu_P(e_Y(X)) - \mu_P(Y)}{\mu_P(Y)} .$$

Von Davier, Holland and Thayer (2004) report that for the first two moments, the PRE is usually small but not zero, and as the moment (*p*) increases, the PRE(*p*) typically increases. They argue that observed score equating cannot convert the discrete

distribution of *X* exactly into *Y*, but kernel equating can work so that the first ten moments of the two distributions are very close.

Step 5: Calculating the SEE and SEED

The final step of kernel equating is the computation of the standard error of equating, or the SEE, which is the standard deviation of the asymptotic distribution of e_Y and the standard error of equating difference, or SEED, which is the difference between two equating functions and can indicate if a linear equating function is sufficient rather than a curvilinear one (von Davier et al., 2004). In the SG design, the SEE is calculated using the variance of the large sample distribution of $\hat{e}_Y(x)$ by:

$$SEE_Y(x) = \hat{\sigma}_Y(x) = \sqrt{Var(\hat{e}_Y(x))}$$

for equating *X* to *Y* and, vice versa:

$$SEE_X(y) = \hat{\sigma}_X(y) = \sqrt{Var(\hat{e}_X(y))}$$

(von Davier et al., 2004). The **C** matrices created in the pre-smoothing step are used here for the calculations. The design function, equating function chosen, and the pre-smoothing of the raw data all play a role. The SEE and SEED are computed differently for the SG and NEAT chain equating (CE) designs. Because CE is computed differently than the other approaches within kernel equating, with ***r*** and ***s*** never actually being calculated, the SEE and SEED are approached differently. More specifically, the design function plays a different role for the NEAT chain equating error calculations.

Typical NEAT kernel equating values for SEE result in a dog-bone shape, higher values occur at the extreme ends of the score scale, and lower in the middle where there

are higher frequencies of examinees. To determine whether the linear equating function is sufficient as opposed to the curvilinear function, the SEED is used. This is done by plotting the differences between the equipercentile equating function and the linear equating function (with four large bandwidths in the NEAT chain equating case, two in the SG case). These differences are then compared to twice the value of the SEED, positive and negative. If a significant portion of the score range falls outside ±2SEED, then the curvilinear model is more appropriate.

Kernel equating has two options under the NEAT design: chain equating and post-stratification equating. Chain equating directly estimates the equating function by using the two single group designs with NEAT, whereas post-stratification, which is the kernel version of the frequency estimation technique, estimates score probabilities on a target population that is a composite of groups *P* and *Q* (von Davier et al., 2004). Chain equating, the older approach of the two, is a two stage process: linking *X* to *A* on *P*, then linking *A* to *Y* using *Q*. Post-stratification equating, which is closely related to the Tucker method (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004), estimates marginal distributions of *X* and *Y* on *T*, the target population, and then equates using those marginal distributions. These marginal distributions are calculated by conditioning on the anchor scores.

Relevant Studies

Comparing IRT methods and traditional observed score methods is not a new concept in equating research. In 1988, Petersen, Cook, and Stocking used Scholastic

Aptitude Test (SAT) data to investigate the differences between IRT equating methods and conventional observed score methods in the NEAT design. Comparing unsmoothed equipercentile, Tucker, Levine equally reliable and unequally reliable, concurrent calibration, fixed *b*'s, and characteristic curve transformation method, the authors found that linear equating methods tend to work better when the tests are reasonably parallel and of equal lengths. When the items and lengths vary, IRT equating using the 3-PL model is more stable, with concurrent calibration being the most stable. The authors found that conventional and IRT methods are equally sufficient, with differing results between the two tests, Verbal and Mathematics.

Similarly, in 1997, Han, Kolen, and Pohlmann explored the differences in IRT true-score equating, observed-score equating, and traditional (unsmoothed) equipercentile equating methods in the NEAT design. Using multiple forms from the Mathematics and English portions of the ACT, the authors compared the results of the three methods to each other and investigated the relationship between the discrepancies in equating results and the difference in difficulty of the two equated test forms. They found that there is a non-significant difference in the equating stability of the two IRT methods, but that both methods are more stable than the traditional equipercentile equating. Han and colleagues conclude that there appears to be a positive relationship between the discrepancies in equating results and the difference in difficulty among the two test forms, and call for further investigation.

In 1998, Kim and Cohen explored three IRT methods of equating and linking in the NEAT design: concurrent calibration based on a posteriori estimation, characteristic

curve transformation method, and concurrent calibration with marginal maximum likelihood estimation. The concurrent methods were calculated with IRT calibration software BILOG and MULTILOG. The authors found that when the anchor test length is short, the characteristic curve method worked better, delivering a smaller root mean square difference (RMSD) than the other methods. However, when the anchor test length was longer (i.e., more than 10 items), the three methods delivered similar results. The IRT model used was a 3-PL and data were simulated.

In 2002, Hanson and Béguin published a report comparing various IRT equating methods under the NEAT design using simulated data. Specifically, they investigated the characteristic curve methods Stocking and Lord (1983) and Haebara (1980), the moment methods mean/mean and mean/sigma, and concurrent calibration in both BILOG-MG and MULTILOG. The authors found that the moment methods (mean/mean and mean/sigma) produced much larger errors than the other methods, and that using BILOG-MG for concurrent calibration produced less error than for separate estimation, but that MULTILOG produced the opposite results.

Hanson and Béguin were not alone in investigating IRT equating methods. Jodoin, Keller and Swaminathan (2003) explored the differences between concurrent calibration, fixed common item parameter estimation (FCIP), and Stocking and Lord's transformation method. Unlike Hanson and Béguin, this used examinee data, and was one of the few that investigated FCIP against the other IRT methods. The authors found that although there was a lot of agreement between the proficiency classifications of the

examinees using the three methods, there was sufficient disagreement to warrant further investigation using simulated data where truth is known.

In recent years, more studies have been conducted investigating the benefits and limitations of kernel equating versus the more traditional methods of observed score test equating. In 2005, von Davier, Holland, Livingston, Casabianca, Grant and Martin used real data in an equivalent groups (EG) design to create pseudo-tests in the NEAT design to compare kernel equating results to those of other more traditional observed score equating techniques. Equipercentile equating and linear equating were investigated under the EG design, and Tucker, Levine observed-score, frequency estimation, and chained equipercentile were studied in the NEAT design, with linear and equipercentile equating both conducted in kernel equating. To do this, the authors created a criterion equating function with the EG designs with which the equating methods in the NEAT design could compare. Two smaller tests of 44 items each were created from a larger test, and an external parallel anchor of 24 items (54.5%) was used. Results indicated that kernel equating approximates an equating that is very close to the more traditional techniques, but that is actually closer to the criterion, and thus more accurate. The effects of differences in test form difficulty, length of anchor test, and sample sizes were not investigated.

In a related study in 2006, von Davier and Ricker studied the role the external anchor test length plays in equating in the NEAT design. Creating a criterion equating using the classical equipercentile method in the EG design, the authors compared the results of kernel equating with large bandwidths (linear equating), optimal bandwidths

(equipercentile equating), and the traditional methods equipercentile and chained equipercentile with external anchor lengths of 24 (54.5%), 20 (45.4%), and 16 (36.4%) items. They established guidelines for comparing results, using a score difference that matters (SDTM), which is any score difference 0.5 or larger. The authors found that kernel equating with optimal bandwidths results are very close to the other equipercentile observed score methods, but that kernel equating with large bandwidths did not closely approximate the other traditional linear methods, especially at the lower end of the score scale, and this method did not come as close to the criterion equating as well as the equipercentile method did. They concluded that the choice of equating function could determine the amount of error in the test scores, especially when the anchor length was shorter.

Similarly, Mao, von Davier, and Rupp (2005) compared kernel equating to traditional methods using PRAXIS data. They found that kernel equating results were very close to the traditional methods, with large bandwidths producing results very similar to the Tucker method, and optimal bandwidths in post-stratification equating very close to frequency estimation. In the EG design, kernel equating results were very close to those of the traditional methods. Mao and colleagues varied the sample size, comparing both the EG and NEAT designs, with both an internal and an external anchor. Like previous studies suggested, their results indicated that kernel equating closely approximates traditional equating methods. With an external anchor, there are differences between kernel equating with optimal bandwidths and frequency estimation are larger at the lower end of the score range, but not meaningfully so. The linear

methods within the EG design were almost identical, a finding which is consistent with previous research. The authors called for more research investigating the accuracy of kernel equating at score ranges where few examinees fall, which is likely when sample sizes are small.

The relationship that sample size plays in kernel equating results has been investigated to some extent. In 2006, Grant, Zhang, Damiano and Lonstein studied small-sample equating with the NEAT design in kernel equating, investigating the effects on the standard error of equating. Previous studies had been conducted with large sample sizes only (over 2000 examinees per form), whereas this study sought to compare the performance of kernel equating when the sample sizes are small: 1000, 500, 250, 125, and 75. With smaller sample sizes, there are more breaks in the score distributions, and results indicated that the equating accuracy increased with the increase of sample sizes, as was to be expected, and that increasing a smaller sample size improved the equating results much greater than increasing a larger sample size. The authors measured equating error using the SEE discussed previously.

In 2006, Holland, von Davier, Sinharay, and Han compared the chain (CE) and post-stratification (PSE) equating methods in kernel equating. In ideal situations, when group differences are minimal, the two methods give the same results. However, when the anchor test indicates that group differences are wider, chain equating is preferable to PSE. Using real data like studies before, the authors created pseudo-tests with the NEAT design from a real test with EG design by ignoring sections of data. The same three anchor lengths were used as created by von Davier and Ricker (2006). Results indicated

that the CE and PSE results were very similar, with CE results slightly closer to the criterion. When the anchor was lengthened, the PSE results consistently improved. The authors concluded that an IRT-based simulation study would eliminate the problem of multidimensionality and content coverage, and could allow the researcher more control over the relationship between the test forms and the anchor.

Summary

Comparing observed score equating methods to those that fall under IRT is not a new concept to measurement and equating research. With the development of the kernel method of equating, however, observed score equipercentile methods are being improved upon. Previous research has established that kernel equating is a sound and stable equating method, oftentimes improving upon the results of traditional methods. However, to date, no simulation studies have been published comparing kernel equating to its more traditional counterparts. The benefits of a simulation study are great: the researcher is allowed full control over the difficulties of the test forms, the ability levels of the examinees, the reliability and length, as well as difficulty of the anchor test, and the relationship of those test forms. This dissertation attempts to remedy this lack of information by creating situations in which truth is known and several equating methods, including kernel equating, are compared and investigated closely for accuracy and applicability to real-life testing situations.

CHAPTER III

METHODOLOGY

This chapter discusses the methods used in this study. It is divided into two sections: one for each study. Both sections describe the data simulation techniques, as well as equating methods and methodological steps.

Study 1

This section describes the steps taken to simulate the data, to pre-smooth the raw score distributions using multiple loglinear models, and to evaluate the models for best fit and impact on the equating results. This will comprise Study One. The second section in this chapter explains the steps taken in Study Two to simulate data, pre-smooth them using loglinear models, and equate using concurrent calibration, Stocking and Lord's transformation method, kernel, and traditional chained equipercentile equating techniques. The reason this chapter is divided into two studies is simple: without knowing the effects of the loglinear model chosen in the pre-smoothing step of kernel equating, it is difficult to compare the kernel equating results to the other methods with the assumption that the results are equally accurate.

The purpose of the first study is to determine the effects of the loglinear model used in pre-smoothing the bivariate score distributions on kernel equating. The first step of the kernel equating method is to pre-smooth using a loglinear model; the impact of the

model chosen on the equating results is, to date, unknown. This study, in alignment with the second study presented below, focuses on the role of the pre-smoothed model in both the Single Group and NEAT designs with the same restrictions and variations within the simulated data.

Three samples of examinees were simulated using ICEDOG (ETS, 2007): 100,000 examinees, 10,000 examinees, and 1000 examinees. These sample sizes were chosen to illustrate the differences in equating results when the number of examinees is quite large, moderate in size, and small. All simulations use the same restrictions on the IRT parameters. For this study, a 2-PL IRT model was used to simulate two test forms with 100 items each. The 2-PL model was chosen and pseudo-guessing (c) parameters set to 0 because of the effect of the c-parameter on the estimation of a- and b-parameters. It is theorized that, with effective instruction, examinees would demonstrate minimal guessing (Hambleton & Swaminathan, 1985). Many researchers prefer the 2-PL model to the 3-PL over uncertainty surrounding the c-parameter and what it truly is measuring (von Davier, personal communication, 2006). For this reason, the c-parameters were set to 0 in all of the data simulations discussed below.

Difficulty (*b*-) parameters for items on forms *X* and *Y* were restricted to be normally distributed between -2.0 and +2.0 with a mean of 0, and with a standard deviation of 0.8. The two forms were simulated to have the same mean discrimination (*a*-) parameter of 0.8 with a standard deviation of 0.2, and were restricted to be uniformly distributed between 0.4 and 1.6. The examinee groups were simulated to have a mean

ability (θ) parameter of 0.0, with a standard deviation of 0.8. Ability parameters were restricted to be between -3.0 and +3.0 and were normally distributed.

Simulated data output for all examinees (*T*) from ICEDOG (ETS, 2006) come in one plain text file, in a space-delimited format where "1" indicates an incorrect response and "2" is a correct response. For LOGLIN's purposes (ETS, 2007), the data must be in separate files for each group, *P* and Q. To separate the data and rewrite them into dichotomous 0-1 format, Fortran programming code was used. Output files produced from the code included dichotomous data files for each examinee group, as well as bivariate score distributions for each group on both test forms. For the SG design, these score distributions are for *X* and Y, respectively. Tables 3.1, 3.2, and 3.3 describe the data for each simulation in the SG design with 100 items, 60 items, and 20 items on each form, respectively.

Table 3.1: Descriptive Statistics: 100 items per form

	X	*Y*	ρ_{XY}
T (N_T= 100,000)	M = 52.39 SD= 18.75	M = 47.79 SD= 21.11	0.96
T (N_T= 10,000)	M = 56.49 SD= 20.42	M = 44.22 SD= 21.03	0.96
T (N_T= 1,000)	M = 53.06 SD= 21.53	M = 46.81 SD= 20.90	0.96

Table 3.2: Descriptive Statistics: 60 Items per Form

	X	*Y*	ρ_{XY}
T (N_T= 100,000)	M = 30.78 SD= 12.84	M = 28.54 SD= 11.54	0.93
T (N_T= 10,000)	M = 32.64 SD= 12.55	M = 28.95 SD= 12.18	0.93
T (N_T= 1,000)	M = 31.96 SD= 12.61	M = 30.59 SD= 13.00	0.95

Table 3.3: Descriptive Statistics: 20 Items per Form

	X	*Y*	ρ_{XY}
T (N_T= 100,000)	M = 9.13 SD= 4.58	M = 9.31 SD= 4.52	0.85
T (N_T= 10,000)	M = 10.79 SD= 4.26	M = 8.31 SD= 4.39	0.81
T (N_T= 1,000)	M = 12.50 SD= 4.69	M = 11.40 SD= 4.45	0.85

The SG design requires that all examinees take both test forms in the same order, form X, then form Y. Due to the nature of the data simulation, test fatigue was not an issue and was not addressed here. This design was later used as criterion equating to which other results from the NEAT design were compared. Therefore, the steps were repeated using the NEAT design to further understand the role that pre-smoothing model plays on the results of kernel equating and can aid in the comparison of the equating results for the second study presented later.

Simulated items were dichotomous, and thusly summed to create raw test scores, ranging from 0 to 100 for a 100-item test, 0 to 60 for a 60-item test, and 0 to 20 for a 20-item test. A bivariate score distribution was calculated using the raw scores, and inputted into LOGLIN for smoothing. Seven models were chosen: 4-4-0, 4-4-1, 4-4-4, 6-6-4, 8-8-4, 10-10-4, and 12-12-4. The first two numbers of the model tell how many moments were preserved from the original score distributions of *X* and Y. The third number indicates the cross-product moments that were preserved. Therefore, the 12-12-4 smoothed model fits the observed data more closely than the 4-4-4, and the 4-4-0 model fails to account for the relationship between the two tests, while the 4-4-1 model only accounts for the correlation between the forms. These seven models were chosen to

further understand the role of the number of moments preserved, and empirical evidence had shown that the differences in the models were close enough to note results from small changes, but large enough to show differences in equating results.

Feedback statistics include Freeman-Tukey residuals, Likelihood ratio chi-square, Pearson chi-square, Freeman-Tukey chi-square, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and the CAIC (Consistent Akaike Information Criterion). These indicate the degree of fit of the smoothed data to the raw observed score distributions. The likelihood ratio chi-square statistic indicates which models fit the raw data distribution. To determine the best-fitting models among those that fit, the Freeman-Tukey residuals were investigated and compared. The smoothed bivariate distributions were then entered into the Kernel Equating software (ETS, 2007) and equated. Results for the best fitting models were then compared to each subsequent model and the equating differences calculated and evaluated.

Study 2

The second study conducted involves comparing kernel equating results to the IRT true-score equating methods, Stocking and Lord (1983) and concurrent calibration (Lord, 1980). Two designs were used in this study: SG and NEAT. Because the Single Group design requires that all examinees take all items in both forms, this allows the creation of a criterion to which the other NEAT equating results were compared. For the NEAT design equatings, examinees were divided into two different groups. One group, *P*, was assigned to take form X. The other group of examinees, *Q*, was assigned to take

form Y. The only items taken by every examinee were those on the external anchor, form *A*.

To simulate the data, data for both test forms and the anchor were created for both examinee groups using a Single Group design, and systematically ignoring blocks of data to emulate a NEAT design (indicated with asterisks in Table 3.4 below). Table 3.4 shows the form by group interaction for one variation. One criterion equating was calculated for each variation of interest, discussed below. With the NEAT design, there are systematically missing data due to the fact that examinees in group *P* only take test form *X* and anchor A, and examinees in group *Q* only take test form *Y* and anchor A. By simulating data in this fashion and equating them as a single group design, that systematic information was no longer missing, and NEAT equating results were compared for proximity.

Table 3.4: Data Simulation Design: Item Blocks

Examinee Group	*X* *Mean* $a_X \approx 0.8$ *Mean* $b_X \approx -0.1$	*Y* *Mean* $a_Y \approx 0.8$ *Mean* $b_Y \approx 0.1$	*A* *Mean* $a_A \approx 0.8$ *Mean* $b_A \approx 0.0$
P Mean $\theta \approx 0.0$	✓	✓*	✓
Q Mean $\theta \approx 0.0$	✓*	✓	✓

There were several variations of interest here. Table 3.5 shows the interaction of the variations in this study for each group difference in ability. Length of total test form was of interest, with varying lengths of 20 items, 60 items, and 100 items for the forms *X* and *Y*. Another variation was anchor length. The anchor items were varied to be 20% of

the total test length, 35%, and 50%. For the 20 item total test length, however, only the 50% anchor was used. The group differences in ability, as measured by θ, between *P* and *Q* were varied to be 0.0, 0.1, 0.2, and 0.4. As simulated in Study 1, the total sample sizes for the NEAT design were 1000, 10,000, and 100,000 examinees in *T*, which, in the NEAT design, is defined by $T = wP + (1 - w)Q$. For the purposes of this study, groups *P* and *Q* were equal in size, so *w* is equal to 0.5.

Table 3.5: Data Simulation: Variations

N_i =	20 items	Anchor =	50%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000
N_i =	60 items	Anchor =	20%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000
		Anchor =	35%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000
		Anchor =	50%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000
N_i =	100 items	Anchor =	20%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000
		Anchor =	35%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000
		Anchor =	50%	N_{Total} =	1000
				N_{Total} =	10,000
				N_{Total} =	100,000

Each dataset was simulated using a 2-parameter logistic model in ICEDOG (ETS, 2007). In addition to simulating the data, ICEDOG was used to perform the Stocking and Lord transformation method on the parameter estimates of the two simulated forms.

To conduct kernel equating, Fortran code was applied to the simulated data from ICEDOG to format as LOGLIN input, as it is in Study 1. Formatted score distributions were then pre-smoothed in LOGLIN, using the best fitting model with the fewest number of parameters, as recommended by Study 1. This resulted in a smoothed bivariate score distribution and a c-matrix, which are then inputted into KE (ETS, 2007). With the exception of the criterion equating, the smoothed distributions were equated under a NEAT chained equipercentile design. The criterion equatings were conducted under the SG design. All equipercentile equatings were restricted to maintain the original score range.

To conduct the concurrent calibration, MULTILOG (Thissen, 2003) was used, due to its ability to allow for different groups. Raw data formatted by Fortran coding was equated, and the item parameters for both forms *X* and *Y*, as well as anchor *A*, for groups *P* and *Q* calibrated in a single run. The resulting item parameter estimates were then used to calculate test characteristic curves using Fortran code, by finding the true score associated with a given θ_i on both form *X* and form *Y*. Thus, there were three steps for this true score equating process (Kolen & Brennan, 2004): (1) a true score for form *X* was identified, (2) the corresponding θ_i for that true score was found, and (3) the true score on form *Y* for that θ_i was found. Because the actual true scores were not known, the observed scores were used.

For the final equating, traditional chained equipercentile equating, Fortran code was used to convert raw scores on form *X* to the scale of the anchor, *A*, and then to the scale of scores on form *Y*. Because this was an observed score equating technique, no conversion to a score scale was necessary. Figure 1 below graphically illustrates the methods taken in this study. Tables 3.6, 3.7, and 3.8 display the means and standard deviations for the simulated raw score distributions within the NEAT design, including the mean and standard deviation for group *P* on test form *X*, the mean and standard deviation for group *Q* on test form *Y*, and the means and standard deviations for each group on the anchor test, *A*.

Figure 1: Chart of Procedures for Study 2

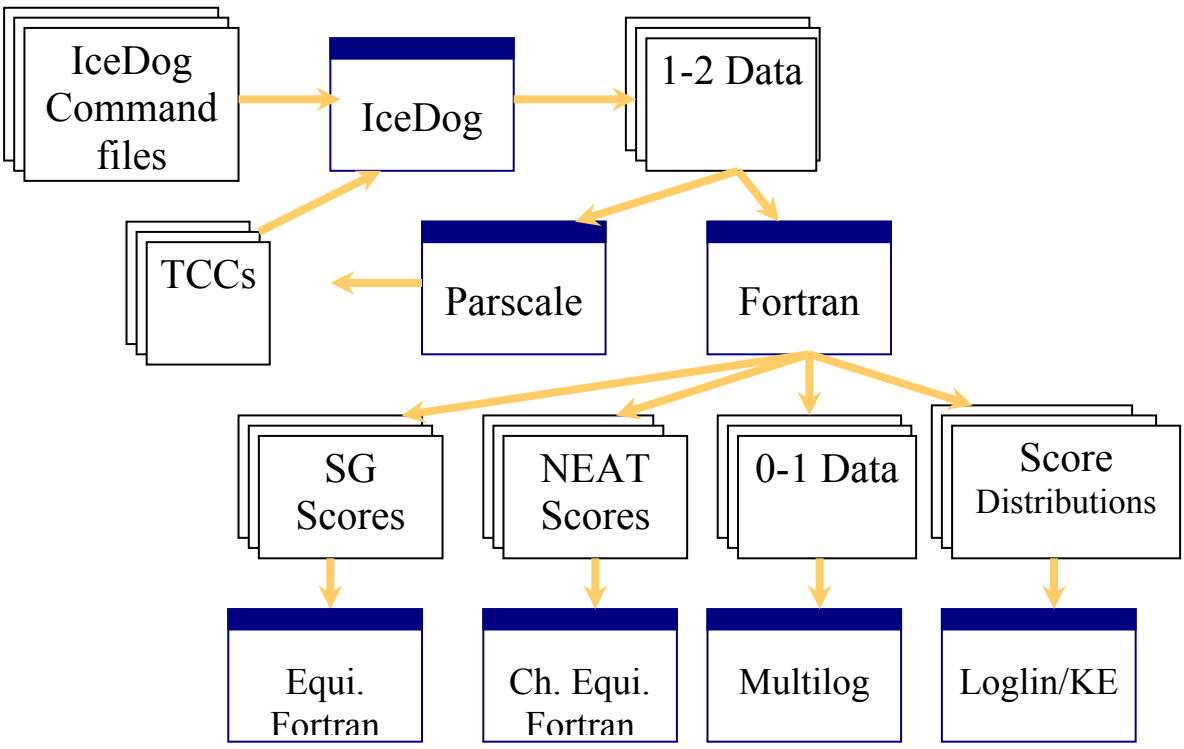


Table 3.6: Descriptive Statistics: 20 Items per Form

Anchor	*Sample Size*	*Difference in Theta*	*Mean X*	*SD X*	*Mean Y*	*SD Y*	*Mean* A_P	*SD* A_P	*Mean* A_Q	*SD* A_Q
50%	1000	0	10.77	4.28	11.08	4.66	5.14	2.38	5.12	2.50
		0.1	12.03	3.79	9.61	4.11	4.63	2.82	4.04	2.83
		0.2	11.03	4.77	9.98	4.24	4.72	2.35	3.96	2.20
		0.4	9.89	4.64	9.27	4.49	5.07	2.38	4.14	2.30
	10,000	0	9.35	4.49	8.03	4.44	4.05	2.51	4.02	2.48
		0.1	10.22	4.09	10.70	3.88	5.06	2.55	4.76	2.52
		0.2	10.26	4.81	9.65	4.56	3.72	2.45	3.20	2.31
		0.4	13.30	3.92	6.87	4.08	4.70	2.20	3.79	2.07
	100,000	0	10.11	4.34	8.92	4.95	5.61	2.16	5.60	2.18
		0.1	12.06	4.04	8.57	4.50	5.15	2.61	4.90	2.61
		0.2	12.28	4.30	8.98	4.31	4.33	2.42	3.81	2.37
		0.4	11.13	4.44	8.64	3.35	4.28	2.35	3.35	2.18

Table 3.7: Descriptive Statistics: 60 Items per Form

Anchor	*Sample Size*	*Difference in Theta*	*Mean X*	*SD X*	*Mean Y*	*SD Y*	*Mean* A_P	*SD* A_P	*Mean* A_Q	*SD* A_Q
20%	1000	0	32.12	13.50	29.06	13.49	6.16	2.72	6.16	2.73
		0.1	33.72	13.36	24.76	12.32	6.51	2.56	6.21	2.61
		0.2	32.75	12.87	30.32	12.39	6.39	3.02	5.81	3.01
		0.4	34.89	11.67	27.58	13.99	8.25	2.53	7.04	2.86
	10,000	0	31.96	12.27	27.66	12.68	6.74	3.04	6.68	3.05
		0.1	31.24	12.29	26.06	11.95	5.49	2.77	5.12	2.75
		0.2	34.97	12.80	27.75	12.27	6.59	2.80	5.93	2.83
		0.4	32.26	13.17	24.33	12.55	6.40	2.81	5.11	2.84
	100,000	0	31.52	12.90	28.94	13.18	7.45	2.66	7.44	2.65
		0.1	32.96	11.73	26.84	11.85	6.65	2.40	6.39	2.43
		0.2	31.22	13.10	28.57	12.76	6.62	2.87	6.01	2.89
		0.4	34.56	12.26	22.51	12.46	7.23	2.71	6.05	2.87
35%	1000	0	30.16	13.41	30.60	13.28	11.50	4.45	11.77	4.51
		0.1	33.59	11.50	30.84	13.58	9.83	4.12	9.33	4.41
		0.2	32.86	13.20	25.54	12.68	10.87	5.03	10.28	5.17
		0.4	32.64	11.77	25.97	11.84	11.44	4.87	9.39	4.74
	10,000	0	29.52	12.19	28.22	12.55	9.56	4.27	9.48	4.23
		0.1	30.72	12.97	27.28	13.43	11.34	4.26	10.87	4.29
		0.2	31.76	12.25	29.52	12.66	11.60	4.29	10.64	4.37
		0.4	33.51	12.46	25.82	12.92	10.83	5.12	8.54	4.87
	100,000	0	30.70	12.88	29.32	12.99	9.43	4.84	9.43	4.84
		0.1	32.04	12.50	29.35	13.16	10.16	4.91	9.55	4.90
		0.2	32.35	12.70	28.31	13.44	10.55	4.31	9.58	4.30
		0.4	33.50	12.92	24.15	12.30	12.97	4.34	10.96	4.48
50%	1000	0	30.99	12.92	29.84	13.00	14.84	6.20	14.69	6.38
		0.1	34.42	13.31	26.76	12.17	13.05	6.47	12.39	6.29
		0.2	32.73	13.90	27.26	12.98	13.61	7.27	12.56	6.90
		0.4	32.54	11.91	28.68	13.06	16.19	6.09	13.10	6.09
	10,000	0	31.16	12.21	26.79	11.74	14.69	6.45	14.73	6.42
		0.1	34.53	12.34	27.92	11.89	15.65	6.94	14.66	6.89
		0.2	32.73	13.05	31.11	12.55	15.18	6.62	13.70	6.44
		0.4	33.96	12.10	24.09	13.17	17.65	6.32	14.73	6.64
	100,000	0	29.04	12.65	28.38	12.82	14.64	6.76	14.58	6.77
		0.1	30.54	12.19	28.24	12.39	15.22	6.79	14.45	6.79
		0.2	33.32	11.70	27.29	11.94	15.30	6.59	13.74	6.58
		0.4	34.28	12.82	25.86	13.36	17.47	6.78	14.20	7.02

Table 3.8: Descriptive Statistics: 100 Items per Form

Anchor	*Sample Size*	*Difference in Theta*	*Mean X*	*SD X*	*Mean Y*	*SD Y*	*Mean* A_P	*SD* A_P	*Mean* A_Q	*SD* A_Q
20%	1000	0	53.52	20.99	50.67	20.68	9.35	4.46	9.35	4.54
		0.1	53.87	19.84	44.12	21.03	9.70	4.50	8.65	4.41
		0.2	54.08	20.41	45.93	21.02	9.96	4.42	8.80	4.45
		0.4	60.61	20.13	42.84	20.33	12.20	4.51	10.39	4.65
	10,000	0	52.37	21.40	48.18	21.41	8.05	4.43	8.14	4.38
		0.1	52.76	20.84	45.65	21.25	10.93	4.69	10.33	4.69
		0.2	54.39	19.92	49.55	21.13	11.20	4.17	10.16	4.28
		0.4	57.30	20.14	42.90	20.29	10.20	4.46	8.13	4.28
	100,000	0	53.72	21.88	46.06	19.83	9.43	4.32	9.43	4.28
		0.1	54.93	19.89	48.49	21.10	10.55	4.71	10.04	4.70
		0.2	54.70	20.60	46.28	19.67	12.29	4.25	11.34	4.36
		0.4	58.63	20.81	41.97	21.22	10.70	4.64	8.59	4.61
35%	1000	0	53.87	22.44	46.86	20.52	16.05	8.24	15.79	7.64
		0.1	52.40	21.56	44.66	19.44	19.38	7.85	18.41	7.61
		0.2	54.72	22.49	46.05	21.77	17.77	7.98	15.87	7.52
		0.4	58.50	19.02	40.27	20.94	18.56	7.23	15.75	7.56
	10,000	0	52.03	20.23	47.57	20.32	17.62	7.21	17.79	7.24
		0.1	49.77	20.99	44.15	20.27	17.47	7.33	16.83	7.43
		0.2	57.92	21.02	44.92	20.50	17.18	7.50	15.62	7.43
		0.4	57.12	20.91	41.97	20.20	17.96	7.55	14.53	7.27
	100,000	0	53.44	21.58	44.97	20.68	15.93	7.77	15.87	7.76
		0.1	54.33	20.86	45.06	19.98	16.88	7.66	15.85	7.61
		0.2	54.02	20.03	42.84	21.08	19.77	7.87	17.89	7.98
		0.4	57.45	20.59	41.19	19.99	19.50	7.41	15.96	7.52
50%	1000	0	51.83	21.68	48.47	19.09	27.54	10.74	28.33	10.46
		0.1	53.06	19.86	46.67	19.79	25.13	10.06	23.58	10.43
		0.2	54.36	20.14	50.06	20.93	26.35	11.02	24.58	11.11
		0.4	55.60	22.33	43.20	20.41	26.76	11.75	21.92	11.12
	10,000	0	52.26	21.28	47.95	21.11	26.27	10.95	26.38	10.95
		0.1	52.03	21.38	45.65	21.93	26.13	10.35	24.43	10.36
		0.2	51.02	20.62	45.51	21.43	23.85	11.03	21.43	10.91
		0.4	57.22	20.52	44.15	20.61	23.43	10.96	18.23	10.38
	100,000	0	49.29	20.11	48.79	21.27	22.35	11.41	22.26	11.34
		0.1	54.42	20.40	44.43	19.94	28.16	10.22	26.78	10.38
		0.2	53.99	20.44	45.38	20.74	27.78	10.94	25.13	11.14
		0.4	57.85	20.56	43.39	21.45	27.04	10.16	22.18	10.16

To compare the results of the four equating techniques studied here, criteria equating functions were calculated. These were completed using a Single Groups design

for traditional equipercentile observed-score equating. Because the data were simulated, there was no concern regarding order effects or fatigue. All examinees were simulated to have taken every item for both test forms and the anchor. The equating function was then compared to the NEAT equating functions calculated by the kernel method, concurrent calibration, and the Stocking and Lord transformation. It was hoped that the kernel equating method was not only close to the other methods studied here, but that its differences from the criterion would be negligible.

CHAPTER IV

RESULTS

This chapter discusses the results obtained from each of the two studies. The first section focuses on Study 1, that investigates the impact of the loglinear model chosen in the pre-smoothing step of kernel equating on the equating results. Seven models were compared. The second section of this chapter focuses on Study 2, that compares the results of multiple equating methods: kernel, chained equipercentile, Stocking and Lord transformation method, and concurrent calibration. These methods were compared to each other and to the criterion equating (i.e., equipercentile equating) under a variety of situations.

Study 1

Chapter 3 discussed the methods used to obtain the data used in this study. Once bivariate score distributions were calculated using the Fortran code, they were input into LOGLIN (ETS, 2007) and the seven models were created. These seven models were then input into KE (ETS, 2007) and the equatings completed. Chi-square fit statistics and Freeman-Tukey residuals were explored to measure fit of each of the models to the original raw score data. The three chi-square fit statistics obtained were Likelihood ratio chi-square, Freeman-Tukey chi-square, and the Pearson chi-square. In all cases, the

4-4-0 model showed poor model fit, with all three chi-square statistics larger than the critical values corresponding to the appropriate degrees of freedom.

Based upon observations of the fit statistics and their relationship with sample size, it appeared that as sample size increased, more moments of the score distributions needed to be preserved. Table 4.1 below shows the minimum (fewest moments maintained) models that showed reasonable fit for each sample size and test form length.

Table 4.1: Simplest Model Fit

Sample Size	*Test Form Length*	*Smallest Fitting Model*
100,000	20 items	6-6-4
	60 items	4-4-4
	100 items	4-4-1
10,000	20 items	4-4-1
	60 items	4-4-1
	100 items	4-4-1
1000	20 items	4-4-1
	60 items	4-4-1
	100 items	4-4-1

Tables 4.2, 4.3, and 4.4 below display the three chi-square fit statistics for each model for each dataset. Table 4.2 shows that the Likelihood Ratio chi-square and Freeman-Tukey chi-square statistics indicated sufficient fit for the 4-4-0 model in the 1000 sample size. However, the Pearson chi-square indicated poor fit, and thus this model was rejected as a viable model to fit the raw observed data. Table 4.3 shows this same phenomenon for the 60-item test length. Table 4.4 shows that the smallest model

that fit for the 100,000 sample size with the 20-item test lengths is the 6-6-4 model. The other sample sizes fit with 4-4-1 models and greater, with agreement among all three chi-square statistics.

Table 4.2: Chi-Square Fit Statistics: 100 Items per Form

		Likelihood Ratio χ^2	*Pearson* χ^2	*Freeman-Tukey* χ^2
N=100,000	**4-4-0**	249366.39*	426247.09*	267409.00*
	4-4-1	5855.76	6132.27	5414.83
	4-4-4	4094.97	4029.53	3630.89
	6-6-4	4063.26	4018.62	3604.02
	8-8-4	4009.43	3961.73	3552.45
	10-10-4	3994.71	3952.97	3538.96
	12-12-4	3967.83	3933.90	3513.78
N=10,000	**4-4-0**	29170.68*	60148.82*	22069.50*
	4-4-1	3290.35	3762.60	2690.73
	4-4-4	3078.24	3416.27	2535.20
	6-6-4	3064.54	3397.73	2528.17
	8-8-4	3076.26	3470.77	2529.50
	10-10-4	3070.94	3489.16	2526.38
	12-12-4	3065.58	3489.39	2521.69
N=1000	**4-4-0**	4701.94	19530.78*	1942.31
	4-4-1	2047.58	3247.16	1278.99
	4-4-4	2017.98	2938.56	1268.74
	6-6-4	2017.18	2893.85	1270.55
	8-8-4	2009.70	2902.05	1267.61
	10-10-4	2001.20	2858.93	1265.30
	12-12-4	1997.34	2821.72	1265.40

*** denotes poor model fit**

Table 4.3: Chi-Square Fit Statistics: 60 Items per Form

		Likelihood Ratio χ^2	*Pearson* χ^2	*Freeman-Tukey* χ^2
N=100,000	**4-4-0**	208062.63*	319798.38*	241421.55*
	4-4-1	3567.71	4226.12*	3454.59
	4-4-4	2073.75	1995.50	1855.02
	6-6-4	1989.55	1861.35	1772.94
	8-8-4	1930.93	1799.07	1714.89
	10-10-4	1900.93	1772.32	1685.87
	12-12-4	1884.90	1757.24	1670.08
N=10,000	**4-4-0**	22195.95*	38329.26*	20465.94*
	4-4-1	1797.79	2841.48	1524.77
	4-4-4	1619.26	1711.16	1374.51
	6-6-4	1612.10	1687.25	1371.66
	8-8-4	1603.22	1694.76	1362.32
	10-10-4	1591.29	1679.56	1352.16
	12-12-4	1589.90	1678.50	1350.94
N=1000	**4-4-0**	3356.85	8784.40*	1834.53
	4-4-1	1100.61	1629.25	785.79
	4-4-4	1067.35	1416.66	773.42
	6-6-4	1060.95	1614.81	770.49
	8-8-4	1060.80	1666.28	769.99
	10-10-4	1058.61	1614.80	769.54
	12-12-4	1058.01	1600.83	769.65

***denotes poor model fit**

Table 4.4: Chi-Square Fit Statistics: 20 Items per Form

		Likelihood Ratio χ^2	*Pearson* χ^2	*Freeman-Tukey* χ^2
N=100,000	**4-4-0**	122892.28*	141589.27*	153251.31*
	4-4-1	1426.98*	1398.65*	1442.37*
	4-4-4	606.29	546.91*	592.71*
	6-6-4	359.52	333.37	341.53
	8-8-4	347.29	321.09	329.60
	10-10-4	341.45	315.35	323.66
	12-12-4	339.98	313.81	322.20
N=10,000	**4-4-0**	11205.69*	12936.38*	12964.59*
	4-4-1	353.26	350.68	330.84
	4-4-4	320.81	307.45	300.46
	6-6-4	304.03	316.95	282.23
	8-8-4	298.26	310.07	277.27
	10-10-4	293.45	304.37	272.53
	12-12-4	287.26	296.97	266.44
N=1000	**4-4-0**	1493.47*	2084.11*	1317.99*
	4-4-1	286.49	369.29	238.42
	4-4-4	245.31	275.74	200.30
	6-6-4	241.71	295.30	196.55
	8-8-4	234.44	328.20	187.61
	10-10-4	233.32	324.90	186.03
	12-12-4	230.97	317.14	184.88

***denotes poor model fit**

In addition to the chi-square fit statistics for each model, the Freeman-Tukey residuals were mapped out for each model in each dataset. These residuals were supposed to follow no specific pattern, and ideally lie between -3.0 and 3.0. Tables 4.5, 4.6 and 4.7 below show the range of Freeman-Tukey residuals for each model with each dataset. Although the 4-4-0 model has already been shown to demonstrate poor model fit for all datasets, and was only included for the purpose of comparison.

Table 4.5: Freeman-Tukey Residual Range: 100 Items per form

		Minimum FT Residual		*Maximum FT Residual*		*Range*	
		X	***Y***	***X***	***Y***	***X***	***Y***
N=100,000	**4-4-0**	-7.12	-12.81	3.84	3.84	10.96	16.65
	4-4-1	-8.20	-12.42	3.98	3.80	12.18	16.22
	4-4-4	-4.12	-5.47	2.32	2.47	6.44	7.94
	6-6-4	-3.58	-5.34	2.33	2.61	5.91	7.95
	8-8-4	-3.20	-4.23	2.38	2.63	5.58	6.86
	10-10-4	-2.77	-2.95	2.33	2.82	5.10	5.77
	12-12-4	-2.15	-3.11	2.61	2.34	4.76	5.45
N=10,000	**4-4-0**	-5.18	-4.28	3.17	2.76	8.35	7.04
	4-4-1	-4.20	-3.26	2.86	2.77	7.06	6.03
	4-4-4	-3.54	-2.46	2.17	2.41	5.71	4.87
	6-6-4	-3.55	-2.48	2.20	2.39	5.75	4.87
	8-8-4	-3.44	-2.61	2.01	2.19	5.45	4.80
	10-10-4	-3.55	-2.59	2.13	2.41	5.68	5.00
	12-12-4	-3.47	-2.42	2.19	2.26	5.66	4.68
N=1000	**4-4-0**	-1.83	-2.22	2.31	2.77	4.14	4.99
	4-4-1	-1.82	-2.10	2.28	2.69	4.10	4.79
	4-4-4	-1.81	-2.00	2.10	3.24	3.91	5.24
	6-6-4	-1.87	-1.99	2.03	3.23	3.90	5.22
	8-8-4	-1.92	-1.92	1.99	3.27	3.91	5.19
	10-10-4	-1.71	-1.90	2.22	3.24	3.93	5.14
	12-12-4	-1.78	-1.98	2.12	3.37	3.90	5.35

Table 4.6: Freeman-Tukey Residual Range: 60 Items per form

		Minimum FT Residual		*Maximum FT Residual*		*Range*	
		X	***Y***	***X***	***Y***	***X***	***Y***
N=100,000	**4-4-0**	-12.93	-8.89	4.93	4.74	17.86	13.63
	4-4-1	-11.87	-11.25	4.79	5.75	16.66	17.00
	4-4-4	-5.04	-3.94	2.78	2.06	7.82	6.00
	6-6-4	-4.21	-3.15	3.13	2.12	7.34	5.27
	8-8-4	-2.82	-2.20	2.93	1.92	5.75	4.12
	10-10-4	-2.18	-1.79	2.11	1.94	4.29	3.73
	12-12-4	-1.72	-1.70	2.55	1.66	4.27	3.36
N=10,000	**4-4-0**	-3.48	-3.39	3.08	2.56	6.56	5.95
	4-4-1	-3.09	-3.66	3.01	2.29	6.10	5.95
	4-4-4	-2.98	-2.40	2.29	2.41	5.27	4.81
	6-6-4	-2.94	-2.56	2.23	2.34	5.17	4.90
	8-8-4	-2.74	-2.45	2.45	2.32	5.19	4.77
	10-10-4	-2.80	-1.98	2.35	2.26	5.15	4.24
	12-12-4	-2.76	-1.80	2.32	2.11	5.08	3.91
N=1000	**4-4-0**	-2.17	-1.61	1.76	1.81	3.93	3.42
	4-4-1	-2.20	-1.57	1.78	1.80	3.98	3.37
	4-4-4	-2.11	-1.55	1.59	1.58	3.70	3.13
	6-6-4	-2.15	-1.53	1.61	1.62	3.76	3.15
	8-8-4	-2.11	-1.50	1.69	1.66	3.80	3.16
	10-10-4	-2.28	-1.33	1.66	1.59	3.94	2.92
	12-12-4	-2.19	-1.46	1.53	1.46	3.72	2.92

Table 4.7: Freeman-Tukey Residual Range: 20 Items per form

		Minimum FT Residual		*Maximum FT Residual*		*Range*	
		X	***Y***	***X***	***Y***	***X***	***Y***
N=100,000	**4-4-0**	-7.13	-7.11	4.44	4.99	11.57	12.10
	4-4-1	-7.48	-7.96	4.69	5.56	12.17	13.52
	4-4-4	-4.77	-3.73	2.82	2.45	7.59	6.18
	6-6-4	-1.64	-1.46	1.46	1.48	3.10	2.94
	8-8-4	-1.88	-1.34	0.91	1.36	2.79	2.70
	10-10-4	-1.53	-0.89	1.02	1.77	2.55	2.66
	12-12-4	-1.38	-0.87	1.00	1.48	2.38	2.35
N=10,000	**4-4-0**	-2.61	-2.06	1.64	2.63	4.25	4.69
	4-4-1	-2.59	-1.99	1.66	2.72	4.25	4.71
	4-4-4	-2.28	-1.45	1.76	2.21	4.04	3.66
	6-6-4	-2.20	-1.74	1.69	1.74	3.89	3.48
	8-8-4	-2.04	-1.43	1.45	1.51	3.49	2.94
	10-10-4	-1.58	-1.11	1.30	1.26	2.88	2.37
	12-12-4	-1.83	-1.15	0.96	1.05	2.79	2.20
N=1000	**4-4-0**	-1.49	-2.30	2.24	1.14	3.73	3.44
	4-4-1	-1.52	-2.23	2.19	1.24	3.71	3.47
	4-4-4	-1.35	-2.31	1.86	1.02	3.21	3.33
	6-6-4	-1.39	-2.19	1.95	0.98	3.34	3.17
	8-8-4	-1.30	-1.53	1.89	1.27	3.19	2.80
	10-10-4	-1.18	-1.49	2.05	1.08	3.23	2.57
	12-12-4	-1.16	-1.37	1.87	1.26	3.03	2.63

The above tables illustrate that the 4-4-0 and the 4-4-1 model were closer to each other in range of Freeman-Tukey residuals and that models showing good fit according to the chi-square statistics had similar ranges in residuals. For further investigation, the residuals were graphed and are included in Appendix A. These graphs revealed that there was more variation in residuals between the seven models when the sample sizes were larger, an observation that held for all test lengths. As test length increased, the Freeman-Tukey residuals appeared to vary more sharply, but this was simply due to the score scale variations for the nine graphs. If all graphs were rescaled to have a 100 score point *x*-

axis, the variations in residuals would appear with all of the datasets. In general, the Freeman-Tukey residuals were smaller with smaller sample sizes. This was a cause for concern when varying sample sizes and comparing model fit, as it appeared that smaller samples had better fit with all models than larger samples, which was not necessarily the case. However, the chi-square statistics did show that more models fit the smaller samples than the large samples. A general rule of thumb, such as "residuals are between -2.0 and 2.0" might not be the best approach to determining best-fitting model if sample size does play a role in the magnitude of these residuals.

The 6-6-4 model was chosen for best overall fit due to its parsimony, chi-square fit statistics across all datasets, and the less extreme Freeman-Tukey residuals. The results of the other six models were then each subtracted from the 6-6-4 results in order to gauge the amount of deviation from a good-fitting model. Graphs of each of the equating functions are displayed in Appendix B, whereas Appendix C shows the differences between each model and the 6-6-4 model in order to highlight the small differences between the models' results.

Equating function results indicate that, as sample size decreases, the differences between the models increases, especially at the extreme ends of the score scales. With the 20-item test length and 100,000 sample size, the two models that deviated most from the rest of the group were the 4-4-0 and the 4-4-1 loglinear models. These two models were shown by the chi-square statistics to have poor fit, and therefore were not considered for practical use in this situation. For the 10,000 sample size, the models follow a similar pattern, although the 4-4-1 model was shown to fit by the chi-square

statistics. The smallest sample size, 1000, had the most variation among the seven models with the 20-item test lengths. However, unlike the larger sample sizes, in this situation, the 8-8-4, 10-10-4 and 12-12-4 models deviated the most from the 6-6-4 model. With the 60-item test lengths, the models' results varied more as the sample size decreased, especially at the extreme ends of the score scale. With the 100-item test lengths, the results were fairly similar the shorter tests: the ends revealed the greatest deviation among the models, and that the models deviated more from each other when the sample size was decreased. With the 10,000 sample size, the models all show fairly similar results, but the 4-4-0 model deviated from the other six models throughout the score scale. Because this model's chi-square statistics do not indicate good model fit, this is not an issue. This model would not be used in practical situations.

The equated scores obtained from the seven loglinear models on each dataset were rounded to the nearest whole number, as would be done in a realistic testing situation. Those values that fell beyond the range of observed Y scores were truncated to be equal to the nearest viable score. Table 4.8 below shows the proportions of scores affected by the model chosen for loglinear smoothing for each sample size and test length variation. For comparison purposes, two proportions were provided: the scores affected by choice among all seven models, and scores that are affected by choice of model between 6-6-4, 8-8-4, 10-10-4, and 12-12-4. These four models were chosen because they demonstrated good fit with all of the datasets.

Table 4.8: Proportions of Model Agreement

	Test length	*Proportion of Equated Score Agreement (all 7 models)*	*Proportion of Equated Score Agreement (largest 4 models)*
N=100,000	*20 items*	0.99	1.00
	60 items	0.83	0.97
	100 items	0.79	0.92
N=10,000	*20 items*	0.85	1.00
	60 items	0.77	0.87
	100 items	0.48	0.82
N=1000	*20 items*	<1.00	<1.00
	60 items	0.91	0.94
	100 items	0.58	0.65

The data represented above are the proportion of equated scores that are uniform across the loglinear models of interest. The first column of data, proportion of score agreement between all models, represents scores that obtain consensus across the seven pre-smoothing models. For instance, with the 60-item tests and 10,000 examinees, 76.69% of equated scores were the same across all seven models. The second column of data, proportion of score agreement between the largest four models, represents proportions of raw scores that equated uniformly across the four best-fitting models: 6-6-4, 8-8-4, 10-10-4, and 12-12-4. When the test lengths are 60-items, and there were 10,000 examinees, the top four models listed above agree on the equated score transformation 86.62% of the time. Scores that fell beyond the reportable range (0 to the maximum scores of 20, 60, or 100) were truncated.

The above table shows that for the large sample size, 100,000 examinees, less than 8% of the scores are affected by the model chosen, among all test lengths, with 100% of equated score agreement when the test length is short (20 items), and nearly 97% of score agreement when the length was 60 items. Of particular interest is the relationship between the length of the test forms and sample sizes on the amount of discrepancies between the loglinear models' results. As sample size was increased, the agreement on equated scores between all models increased, and as test length was decreased, the agreement between models increased. With large sample sizes (10,000 and 100,000), the 20-item test has greatest agreement, with all four largest loglinear models reaching the same equated results. With a small sample and large number of items, the models produce the most discrepancy between results, only agreeing on 58% of scores, with the top four models agreeing on 65% of scores. In favorable situations, such as large numbers of examinees, the amount of agreement between models was quite large.

To further illustrate the differences in the equating functions produced by each of the loglinear pre-smoothing models, each model's results were subtracted from the results of the 6-6-4 model and graphed. Figures C.1 through C.9 in Appendix C display these differences. Most discrepancies happen at the high and low ends of the score scale, and are larger when sample sizes are smaller. The 4-4-0 model disagrees with the other models the most frequently. Because it does not take into account the relationship between the two test forms, the smoothed distributions maintain zero cross-product moments of the original raw score distributions. This result was to be expected. The

greatest differences among all seven models occurred when the sample size was 1000 and the test length was 100 items per form, which was also demonstrated in Table 4.8, that stated that the models agree on only 58.2% of scores. On the other hand, graphs C.7, C.8, and C.9 showed that the level of agreement was high among the models, especially the four models with the greatest number of preserved moments.

The standard errors of equating (SEEs) are graphed in Appendix D. It is important to note that in order to show each model's SEE and highlight differences between each model, the scales are not uniform. The graphs indicate that as the length of the test forms decreased, the SEEs decreased overall (holding the sample size constant). As sample size decreased, the SEE increased, and the differences between models increased. The largest standard errors of equating were obtained in the condition with the small sample size (1000 examinees) with the largest test length (100 items), with some values exceeding 2.0. Across the majority of the score scale, the worst-fitting 4-4-0 model obtained the highest standard errors of equating in all instances.

Study 2

This study compared the equating techniques of kernel, traditional chained equipercentile, concurrent calibration, and Stocking and Lord transformation method. Overall, the four comparison methods produced fairly similar results, with some discrepancies at the extreme ends of the score scales. Equating functions for all five techniques are shown in Appendix E. Concurrent calibration, Stocking and Lord transformation method (TBLT), kernel equating, and chained equipercentile equating are all conducted using the NEAT design, whereas the criterion equating, equipercentile

equating, used the SG design, and was shown as a reference point to which the other methods are compared. The figures in Appendix F are included to demonstrate the differences between each of the equating methods of interest and the criterion equating. These figures show much more clearly the differences produced by each of the methods.

Figures F.1 through F.4 display the 100 items-per-form tests with 50 anchor items and 1000 examinees total. All four methods show deviations from the criterion at the ends of the score scale, with the true score methods, concurrent calibration and TBLT, showing the most extreme deviations occurred when the group differences were larger. When group differences are minimal, the methods produced very similar results, with chained equipercentile equating performing best at the extreme ends of the scale. As group ability differences increased, chained equipercentile equating results deviated more from the criterion, with TBLT results indicating more accurate performance, and kernel results very similar.

Figures F.5 through F.8 show the differences in results when the tests are 100 items-per-form with 50 anchor items and 10,000 examinees. When group differences were null, the kernel and TBLT results were very similar, along with chained equipercentile, while concurrent calibration tends to overestimate the equated scores. When group ability differences increased, the methods tended to deviate more at the ends of the score scale, with kernel and chained equipercentile equating deviating the most at the extreme ends. Concurrent calibration, on the other hand, deviated more from the criterion method when group ability differences were increased.

Figures F.9 through F.12 show the equating results differences when the tests were 100 items-per-form, with 50 anchor items, and 100,000 examinees. Concurrent calibration results deviated the most from the criterion equating results across the group ability difference variations, however the other methods were very similar to each other across a majority of the score scale. There was a decrease in similarity as the group differences increased.

Figures F.13 through F.16 show the results of equating on data where there are 100 items-per-test form, 1000 total examinees, and only 35 items on the external anchor test. The graphs show that all four methods of interest give different results across the score scale, with the kernel and TBLT equating results deviating similarly from the criterion when groups had approximately the same mean ability. Chained equipercentile equating performs closely to TBLT and kernel equating, but deviated a bit more at times. Concurrent calibration, on the other hand, deviated the most and overestimated scores on the top half of the score scale. The largest deviations occurred at the extreme ends of the score scale. When group ability differences increased, all methods tended to deviate more from the criterion, just as they did when the anchor test length is 50 items.

Figures F.17 through F.24 show the equating results on the same type of data as above, but with larger numbers of examinees: 10,000 and 100,000 examinees, respectively. When the sample size is 100,000, chained equipercentile, TBLT, and kernel equating results were very similar to each other, despite magnitude of group ability differences. Concurrent calibration continued to deviate the most from the criterion

equating, and the extreme ends of the score scale continued to be the location of most discrepancies between equating results.

The smallest anchor length investigated in this study was 20% of the total test length. When the test forms were 100 items-per-form, the anchor is 20 items long. Figures F.25 through F.36 display this anchor length, varying sample sizes and examinee mean group ability differences just as above. Holding group ability differences constant at zero, Figure F.33 illustrates that the methods were much more similar to each other, with the exception of concurrent calibration, when the sample size was very large at 100,000. Figure F.29 shows the same type of test situation, but with 10,000 examinees, and the equating results deviate by up to half of a score point throughout most of the score scale. Concurrent calibration with the largest sample size revealed an overestimation of equated scores, by up to four score points, through a large portion of the score scale. When the sample size was a bit smaller at 10,000 examinees, concurrent calibration actually produced more similar results to the other methods, but still overestimated the equated scores by up to 1.5 score points through a large portion of the score scale. As group ability differences began to increase and the sample size was 100,000, TBLT performed very close to the criterion, with negligible differences, whereas kernel and chained equipercentile equating results began to deviate more from the criterion by up to a point in the middle of the score scale and up to three points at the extreme end of the scale. Concurrent calibration continued to deviate by up to four score points across the score scale.

Figures F.37 through F.40 explore the equating results when there were 60 items per-test-form. The anchor lengths of 50%, 35%, and 20% (or 30 items, 21 items, and 12 items) were varied just as they were with the 100-item tests, and the sample sizes varied with 1000, 10,000, and 100,000. With this shorter test, and a large sample size of 100,000 examinees with an anchor test of 30 items, TBLT did not perform as closely to the criterion equating as it did with a longer test, deviating up to one score point when groups were approximately equal in mean ability, and up to four score points as group ability differences increased. Kernel equating continued to give results very close to the criterion, only deviating more than half of a point at the extreme ends of the score scale. Likewise, chained equipercentile equating results were very close to the criterion, with any differences being negligible. Concurrent calibration continued to deviate from the criterion equating across the score scale by up to two score points. These observations held true for the smaller anchor length of 21 items as well, with more differences among the methods when group differences grew. However, TBLT performed better when the anchor length was 21 items than when it was 30 items. With the shortest test length, the trend continued, with the differences between the methods being slightly more pronounced, and TBLT performing even closer to the criterion. All methods continued to deviate from the criterion at the extreme ends of the score scale. Reducing the sample size increased the differences between the methods, as it did with the 100-item test length variations.

The smallest test length, 20 items, was treated slightly different than the other two test length variations. Because the number of items was small, only a 50% anchor length,

or ten items, was investigated. Anything less than ten items appeared to lead to a low reliability and thus poor equating in a practical situation, and so it was avoided here. In this case, concurrent calibration continued to deviate from the other methods and the criterion by up to four score points. Kernel equating and TBLT results were quite favorable, often deviating less than half of a score point. At the extreme ends of the score scale, kernel equating results were often up to one score point different than the criterion, but this was not always the case (i.e., Figure F.84). With the 100,000 sample size, chained equipercentile equating results are almost exactly the same as the criterion results, but the differences grew as the sample size was reduced. When group ability differences were increased, kernel, chained equipercentile, and TBLT results deviated more from the criterion, and were more pronounced as the sample sizes was decreased. In the case of 20-item tests, no method deviated more than four points (concurrent calibration), and the observed score equating techniques did not deviate by more than two points (see Figure F.80).

Results indicated that when sample size was large, the equating techniques tended to give very similar output. Stocking and Lord's transformation method was affected by the anchor length when the test length is shorter (60 items), as it was using fewer total items to calculate the test characteristic curves for the transformations. Kernel equating, although it showed some deviation from the criterion at the ends of the score scale, frequently provided a stable and accurate equating function, often deviating less than half of a score point from the criterion, across sample sizes, group differences, and anchor lengths. In situations when kernel equating results showed variation from the criterion,

the other methods showed similar behavior. Concurrent calibration results suggested that it is an undesirable equating technique as compared to the other methods. Chained equipercentile equating oftentimes produced results that were quite similar to the criterion, with the exception of when the sample sizes were small (1000 examinees). Depending on the design of the test and use of the scores, the equating technique chosen could potentially play a key role in the decisions made based on the equated score results. With this in mind, the equating technique should be chosen accordingly.

CHAPTER V

DISCUSSION

This chapter discusses the implications, limitations, and future directions of the research presented here. The first part of this chapter discusses the implications of Study 1, as well as directions for future research. The second part of this chapter discusses Study 2 implications and limitations to the research.

Study 1

The main purpose of this study was to investigate the impact of the loglinear model used in pre-smoothing on kernel equating results. When pre-smoothing, the kernel equating software (ETS, 2007) provided fit feedback such as chi-square statistics and Freeman-Tukey residuals. This information could assist the user in making a decision regarding the loglinear model chosen in this step. However, if the model chosen makes little to no difference on the equating results, then the process of choosing a model may be extraneous and inefficient. Seven models were chosen based on empirical observations of their differences on various datasets. These models ranged from 4-4-0, that only preserved the first four moments of the two test forms and ignored any relationship between the two, and 4-4-1, that preserved the first four moments of the two test forms score distributions and maintained only the correlation between the two forms, to the 12-12-4 model, which preserved the first 12 moments of each test form, as well as

four cross-product moments between them. Larger sample sizes required more moments to be preserved in order to demonstrate adequate model fit.

Findings indicate that, in general, the 6-6-4 model was sufficient for pre-smoothing. It showed adequate fit with every data variation, and the Freeman-Tukey residuals were acceptable in range. The only situation in which a large number of examinees were impacted by which well-fitting model was used for pre-smoothing was with a small sample size and a long test form. In this case, the frequencies for each score possibility were relatively small, making this an undesirable situation for practical testing purposes. When the test length was small (20 items) and the sample sizes were large, (10,000 or larger), no examinee's score was affected by the loglinear model chosen to pre-smooth the raw score distribution.

In addition to similar equated score results, the 6-6-4, 8-8-4, 10-10-4, and 12-12-4 models demonstrated very similar standard errors of equating. These increased overall when test form lengths increased or sample size decreased, with sample size playing a greater role in determining the magnitude of the SEEs. The worse fitting models, the 4-4-0 and 4-4-1, generally obtained higher SEEs across the score scale in all data variations. This information, combined with the lack of consistent evidence for adequate fit from the chi-square statistics and the large Freeman-Tukey residuals made these two models less desirable candidates for pre-smoothing, whereas the 6-6-4 model appeared to be an acceptable choice for trustworthy results in all variations explored here.

Recommendations from this study were as follows: use a loglinear model that fits the data according to the three chi-square fit statistics; use the Freeman-Tukey residuals

to understand the stability of the models and to determine if one well-fitting model is more desirable over a simpler but still well-fitting model; and, in most situations, it is sufficient to use the most parsimonious model. That is to say, the model fitting the fewest number of moments with acceptable chi-square fit statistics and Freeman-Tukey residuals is oftentimes sufficient, and over-fitting the data is not necessary.

A possible extension of this research for future study is the growing use of pre-smoothing in traditional equipercentile equating. Loglinear models are not only useful in kernel equating, but also possibly for other equipercentile observed score equatings. Because the data presented here were simulated, they represented best possible scenarios, and were generally cleaner with no omissions, mistakes, and so on. The next step for this study would be to conduct it on real data from an actual testing program to highlight the differences in results based on the model chosen, and using different design functions, such as the NEAT design, which is a widely-used design.

Study 2

The purpose of this study was to investigate the performance of the kernel method of test equating on simulated data as opposed to other more traditional counterparts: chained equipercentile, Stocking and Lord transformation method, and concurrent calibration. Data were simulated using a 2-parameter logistic IRT model and both examinee and item parameters were constrained to fit a variety of situations. Overall, kernel equating results were very similar to the other methods of test equating, and oftentimes produced very close results with those of the criterion equating.

Results indicated that kernel equating is a viable option for an operational equating technique, especially in situations where true score equating may not be the most stable approach (i.e., small sample sizes or fewer items). Concurrent calibration, like previous research has suggested, oftentimes produces inaccurate results that could have a potentially large impact on test-takers, depending on the purpose and use of the test scores. Stocking and Lord's transformation method produced typically favorable results. However, this may have been an artifact of the nature of the simulation (i.e., the data were generated using an IRT model) and not the quality of the technique itself. In order to test this further, real data should be used, equating using both traditional unsmoothed equipercentile equating and TBLT. One may want to also compare kernel results at that time, to further understand the relationship between the two approaches and the benefits and weaknesses of using them.

Limitations of this study include, as mentioned above, the nature of the simulated data. Because they were created using item response theory, the true score equating techniques used here have an advantage over the observed score techniques. In addition, because the item parameters were simulated, the possibility of obtaining unrealistic or unpredictable score distributions is cause for concern. In several situations, the score distributions were unexpected and equating results rather challenging. With simulated data, this rarely happens, and should be explored further using real data such that results are more realistic and applicable to specific testing programs.

Another limitation to this study was that of Multilog, which can read a maximum of 99,999 score patterns. Thus, the largest sample size had to be broken down into two

samples (one which contained examinees 1-99,999 and the other contained examinees 2-100,000). Both samples were run, and results compared to make sure they were equal. This did not produce any differences greater than 0.002, and parameter estimates from the first "sample" of examinees were used.

Future research stemming from the methods described here includes using real data to obtain more realistic results for practical conclusions, and using alternate methods of data simulation to remove the advantage of the IRT equating methods. Another extension would be to create unequal sample sizes and investigate the role that they play in the equating results.

As the studies presented here show, equating results oftentimes depend on the method chosen. No one method is consistently and universally more accurate than the others, but several are shown to be overall more stable and typically more accurate than others. Kernel equating, the primary method of interest here, was shown to be stable and accurate in most situations, although frequently deviating from the criterion at the extreme ends of the score scale. Fortunately, most of the time, few to no scores were affected by this, as very few examinees fall within these extreme scores. With increasing dependence on test scores for decisions regarding admissions, licensure, placement, and certification, the results of any and all equatings must be as accurate as possible. With the variety of methods and approaches to choose from, it is important to understand the strengths and weaknesses of each method, and to choose carefully and according to needs and challenges of the testing program.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Tests.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Braun, H. I. & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test Equating* (pp. 9-49). New York: Academic.

Cook, L.L. & Eignor, D. R. (1991). IRT Equating Methods. *Instructional Topics in Educational Measurement.* NCME.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.

Grant, M. C., Zhang, L., Damiano, M. & Lonstein, L. (2006). An evaluation of the kernel equating method: Small sample equating in non-equivalent groups. *Paper presented at the national conference of AERA/NCME, 2006.*

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22,* 144-149.

Hambleton, R. K., Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park: Sage Publications, Inc.

Han, T., Kolen, M. & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 105-121.

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item nonequivalent groups equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Holland, P. W. & Thayer, D. T. (1981). *Section pre-equating: The Graduate Record Examination.* Program Statistics Research Technical Report No. 81-13, Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*(2), 133-183.

Holland, P. W., von Davier, A. A., Sinharay, S. & Han, N. (2006). Testing the untestable assumptions of the chain and post-stratification equating methods for the NEAT design. *Research report ETS, 2006*.

Kim, S. H. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating Methods and Practices*. New York: Springer-Verlag.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement*, *18*, 109-118.

Liu, J. & Low, A. C. (2006). An exploration of kernel equating using SAT data: Equating to a similar population and to a distant population. *Paper presented at the national conference of AERA/NCME in 2006, San Francisco, CA.*

Livingston, S. A., Dorans, N. J. & Wright, N. K. (1990). What combination of sampling and equating methods work best? *Applied Measurement in Education*, *3*, 73-95.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum Associates, Inc.

Mao, X., von Davier, A. A. & Rupp, S. (2005). *Comparisons of the kernel equating method with the traditional equating methods on PRAXIS data* (ETS Research Report). Princeton, NJ: Educational Testing Service.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

Ricker, K. L. & von Davier, A. A. (2006). The role of the anchor test in a non-equivalent groups design. *Paper presented at the National Conference of Measurement in Education national conference, 2006.*

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Thissen, D. M. (2003). MULTILOG for Windows (version 7.0.2327.3). Scientific Software International, Inc.

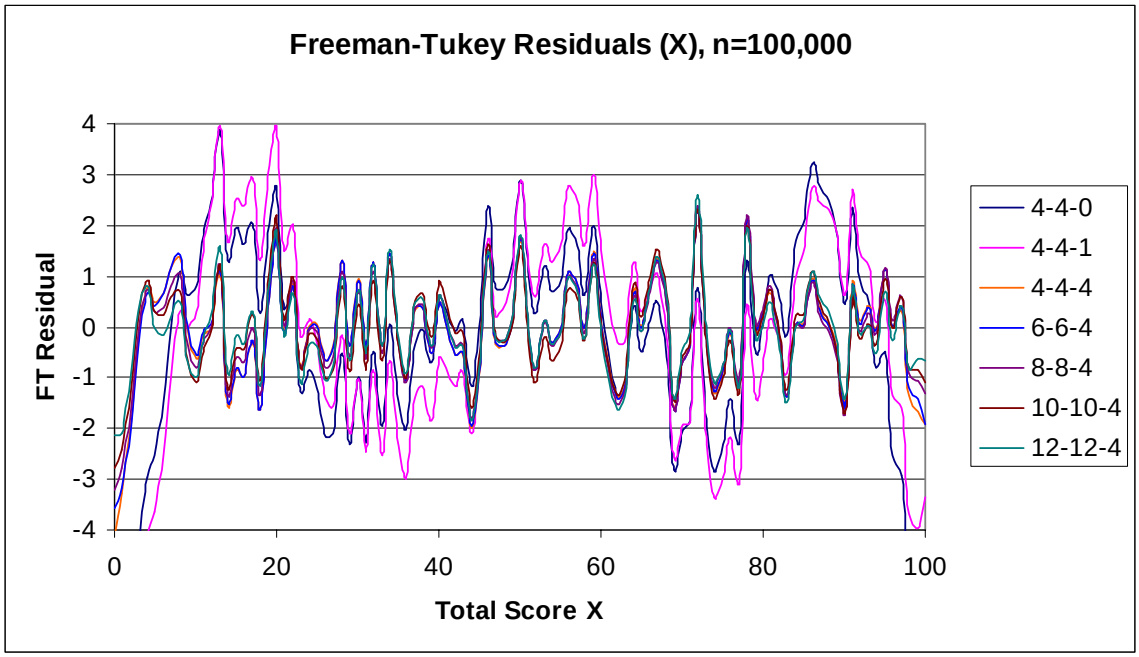
von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). An evaluation of the kernel equating method in a non-equivalent groups design with an external anchor: A special study with pseudo-tests from real test data. *AERA/NCME paper presented in 2006.*

von Davier, A. A., Holland, P. W., Thayer, D. T. (2004). *The Kernel Method of Test Equating*. New York: Springer-Verlag.

von Davier, A. A. & Ricker, K. L. (2006). The role of the anchor test in a non-equivalent groups design. Unpublished research report.

APPENDIX A: Freeman-Tukey Residuals

Figure A.1
100 Items per form, n=100,000



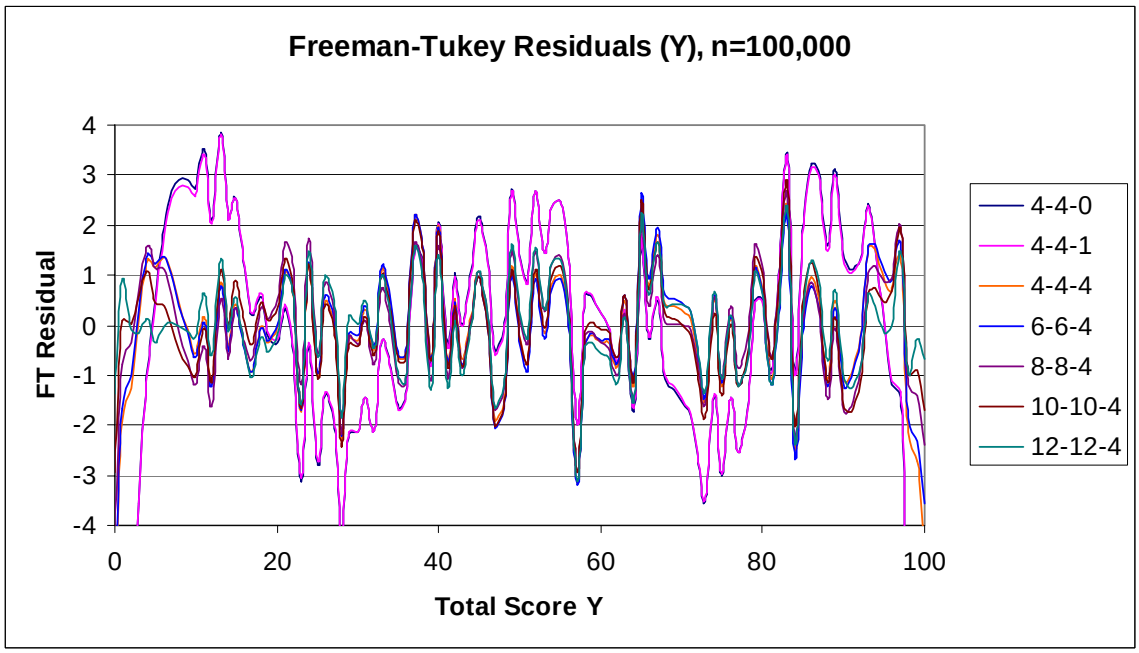
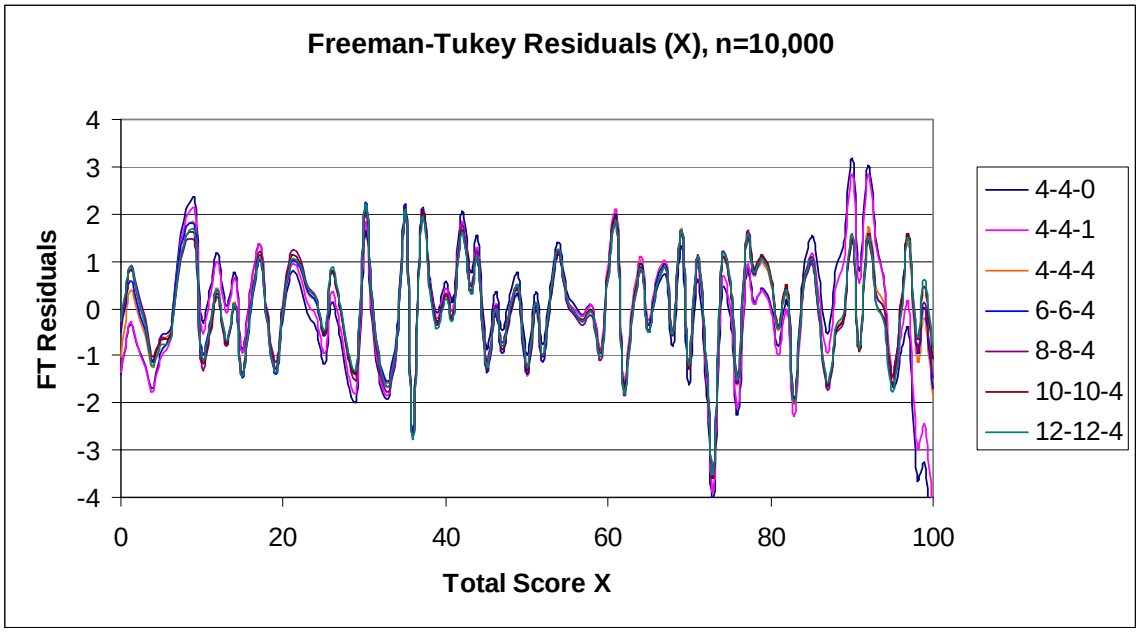


Figure A.2
100 Items per form, n=10,000



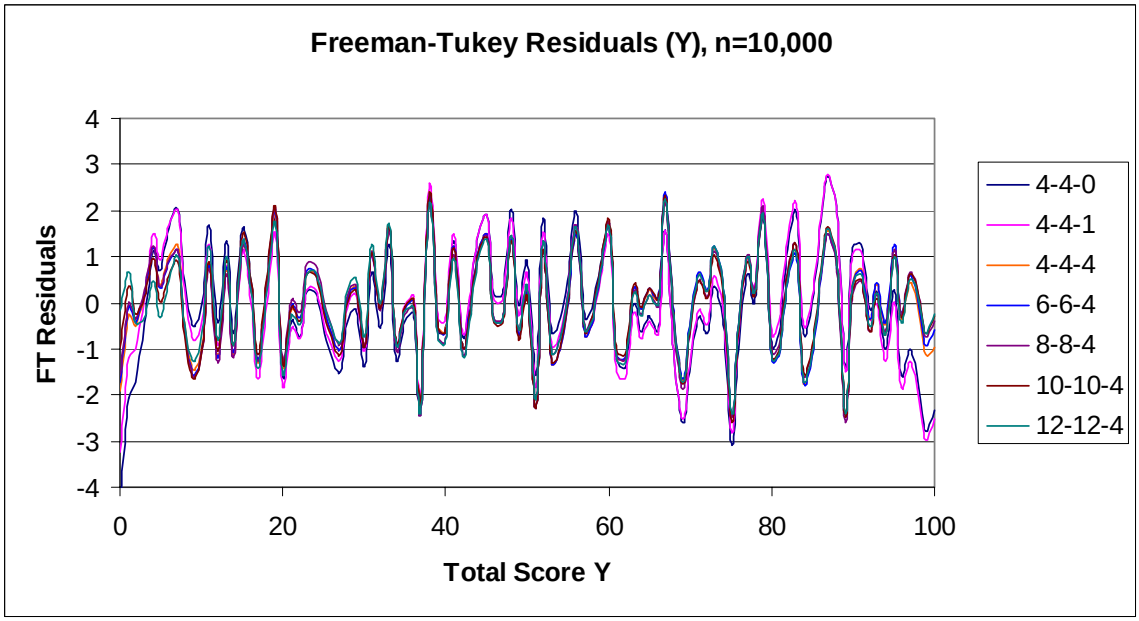
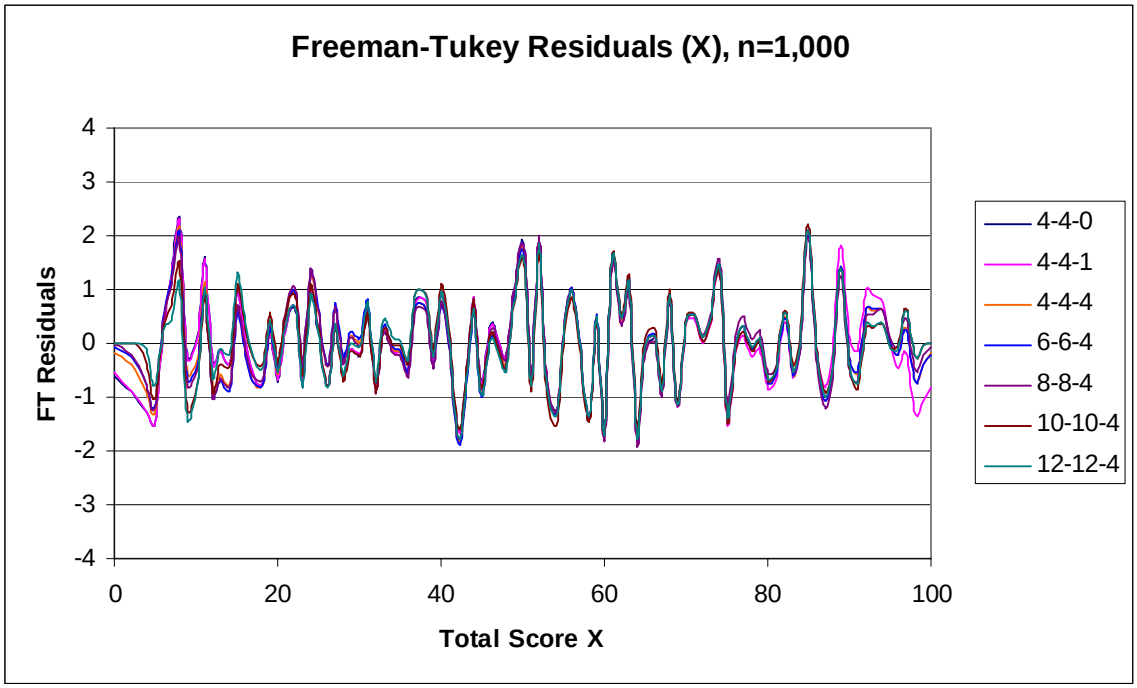


Figure A.3
100 Items per form, n=1000



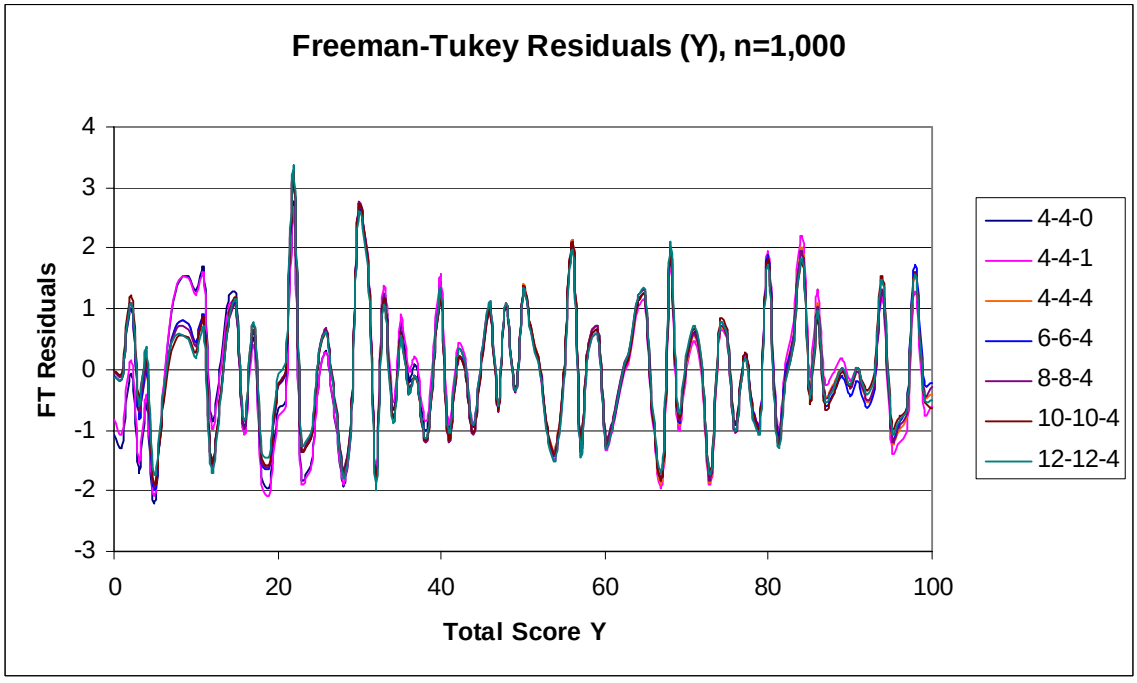
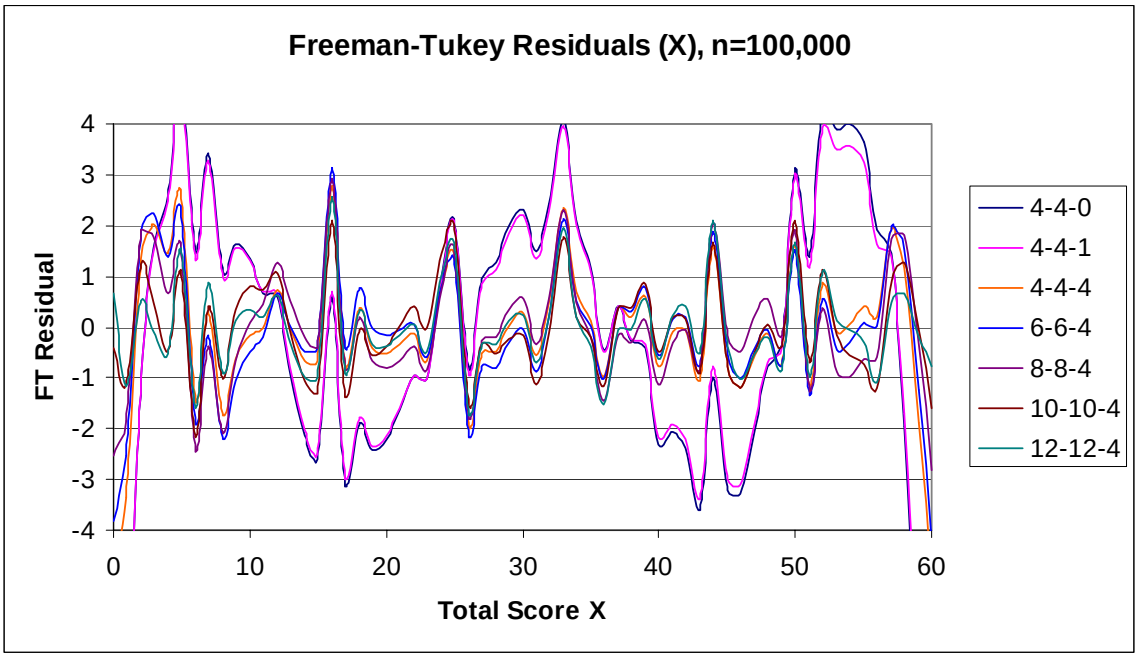


Figure A.4
60 Items per form, n=100,000



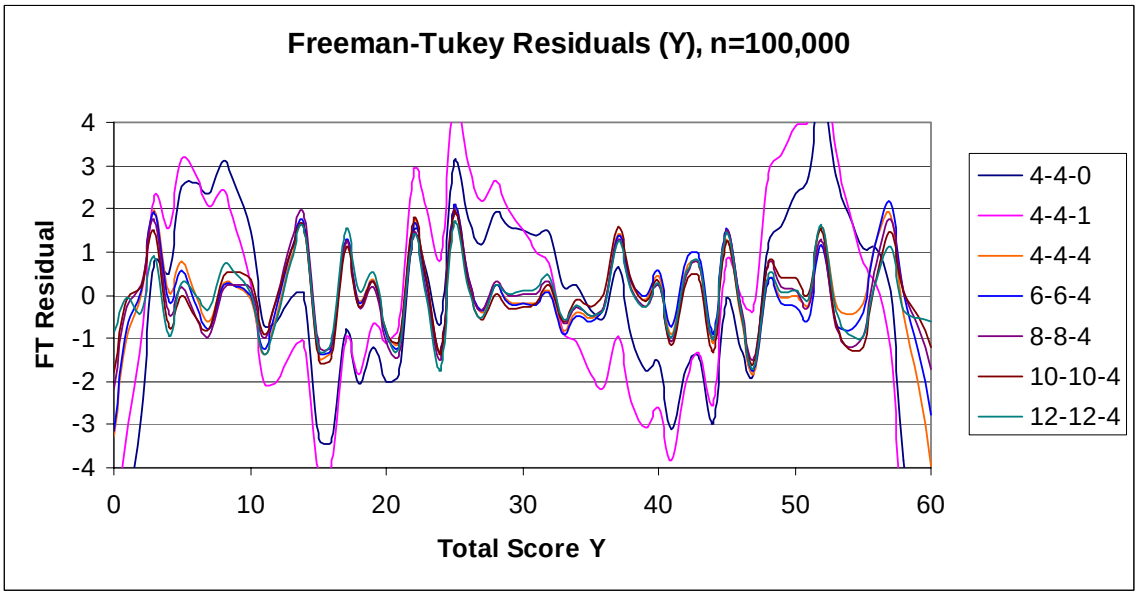
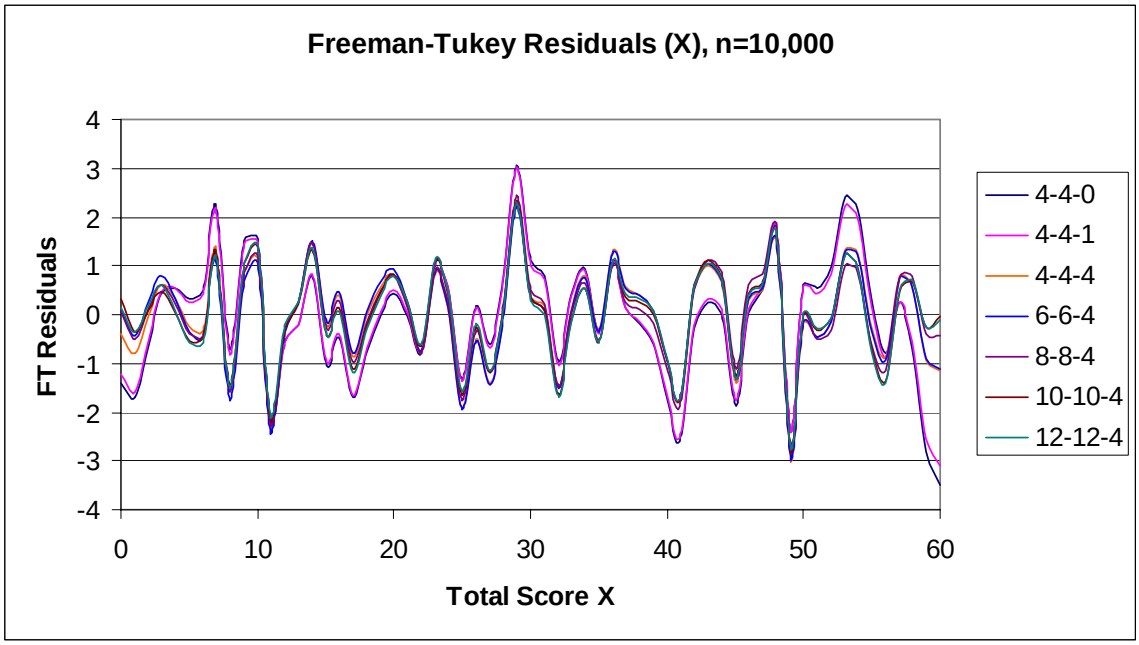


Figure A.5
60 Items per form, n=10,000



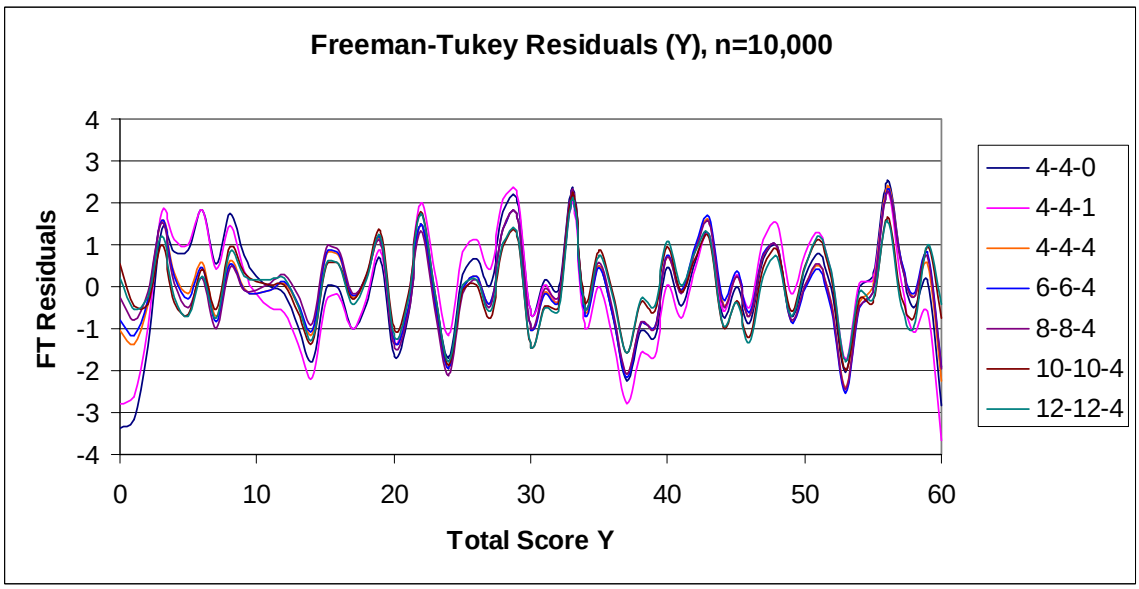
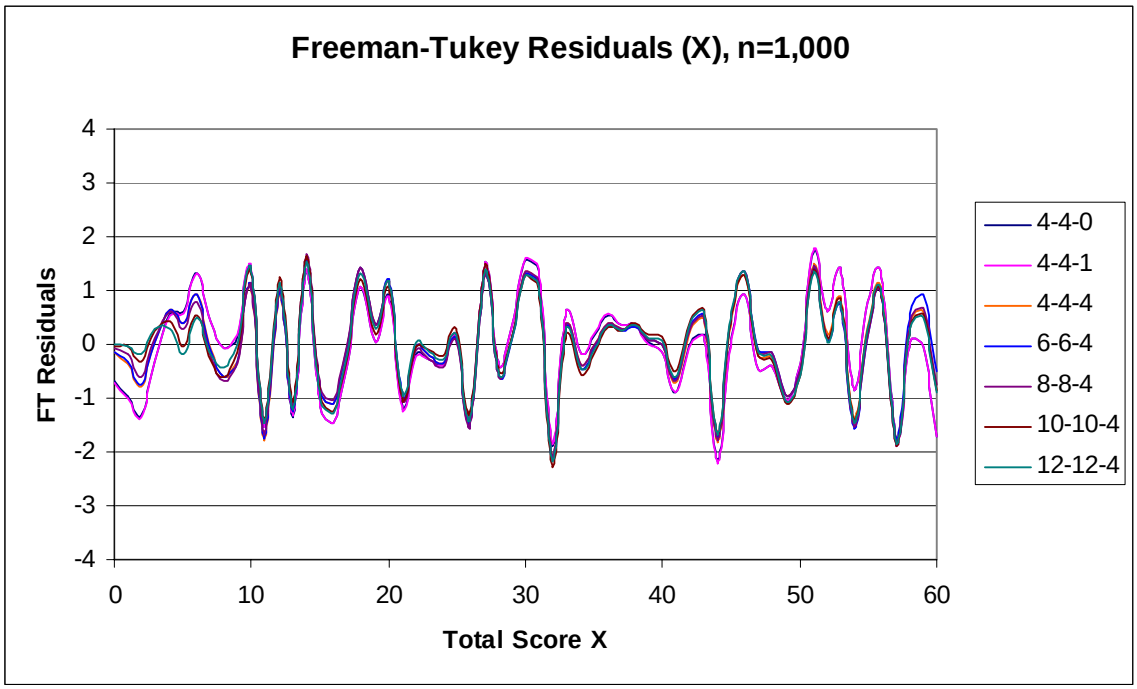


Figure A.6
60 Items per form, n=1000



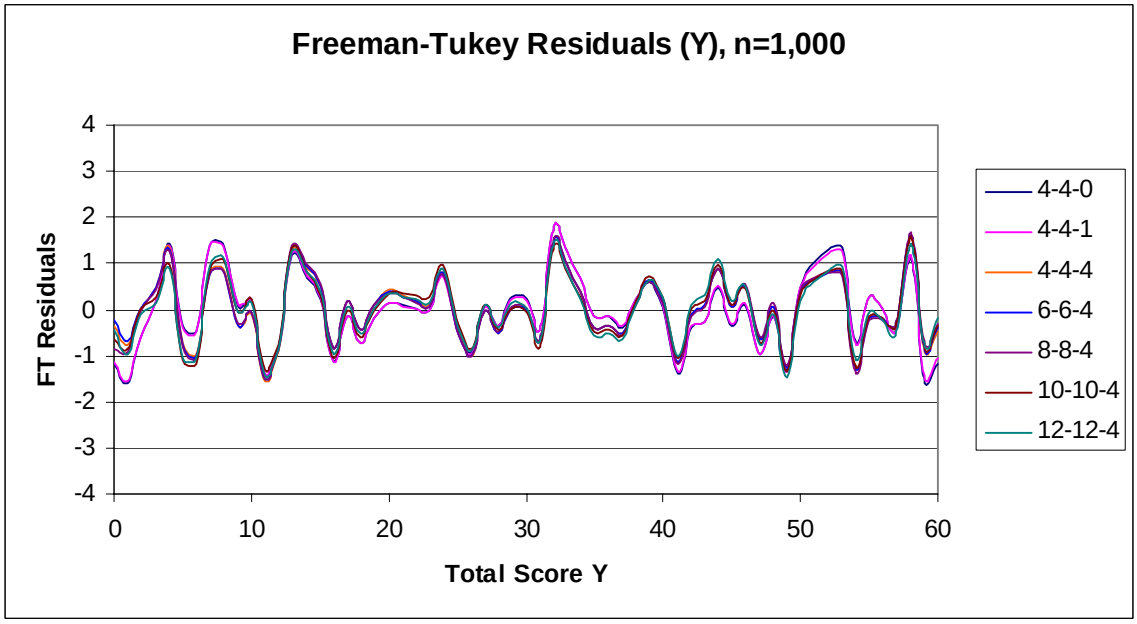
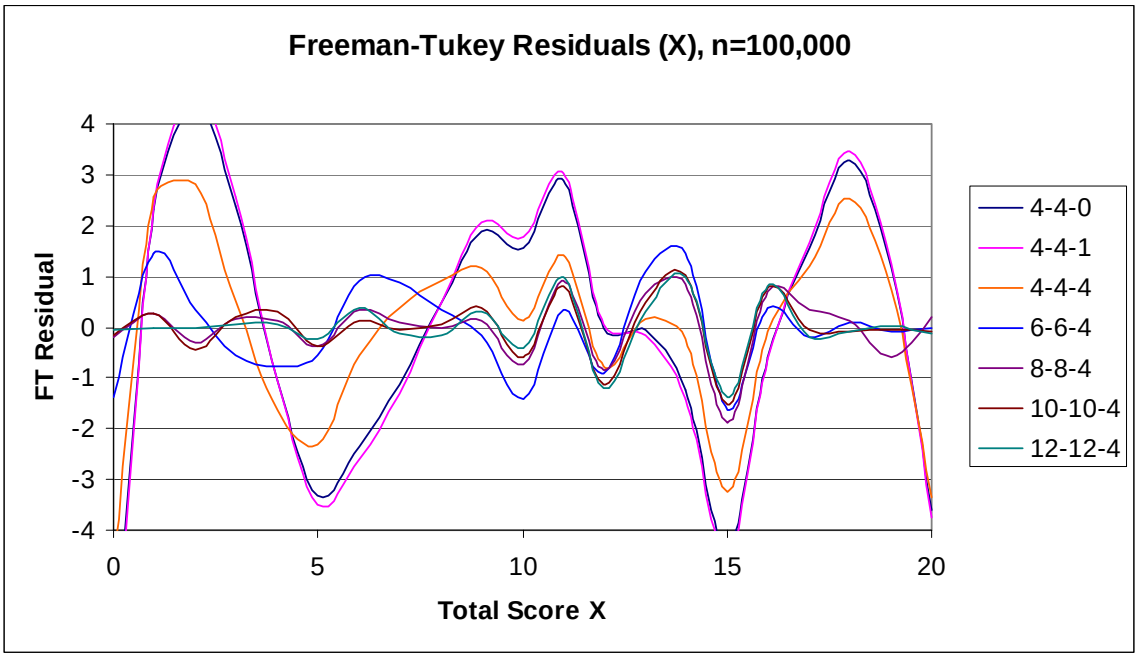


Figure A.7
20 Items per form, n=100,000



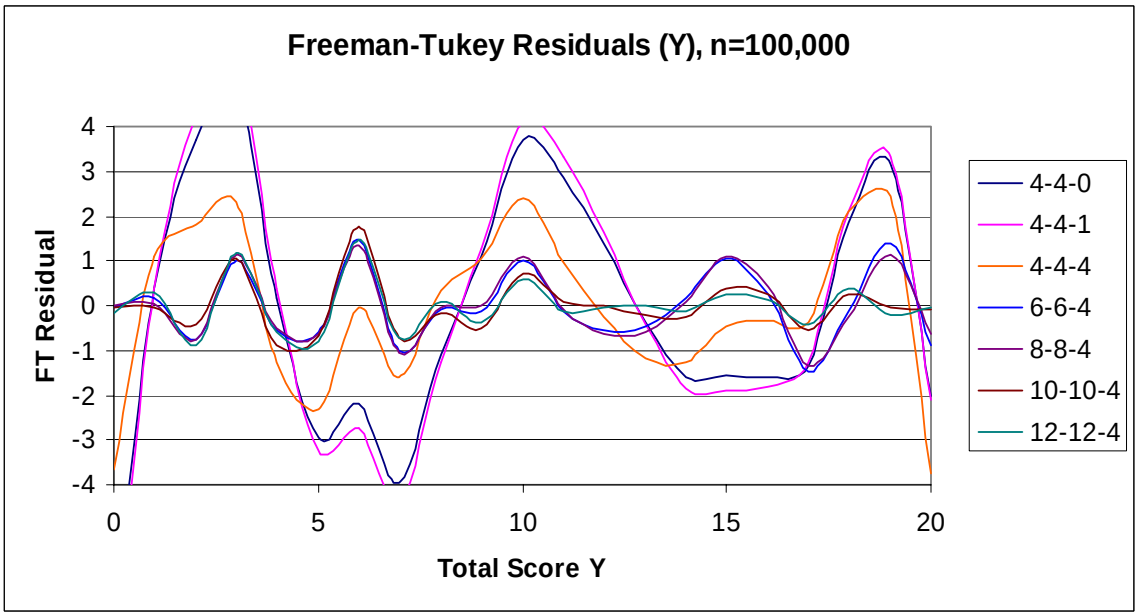
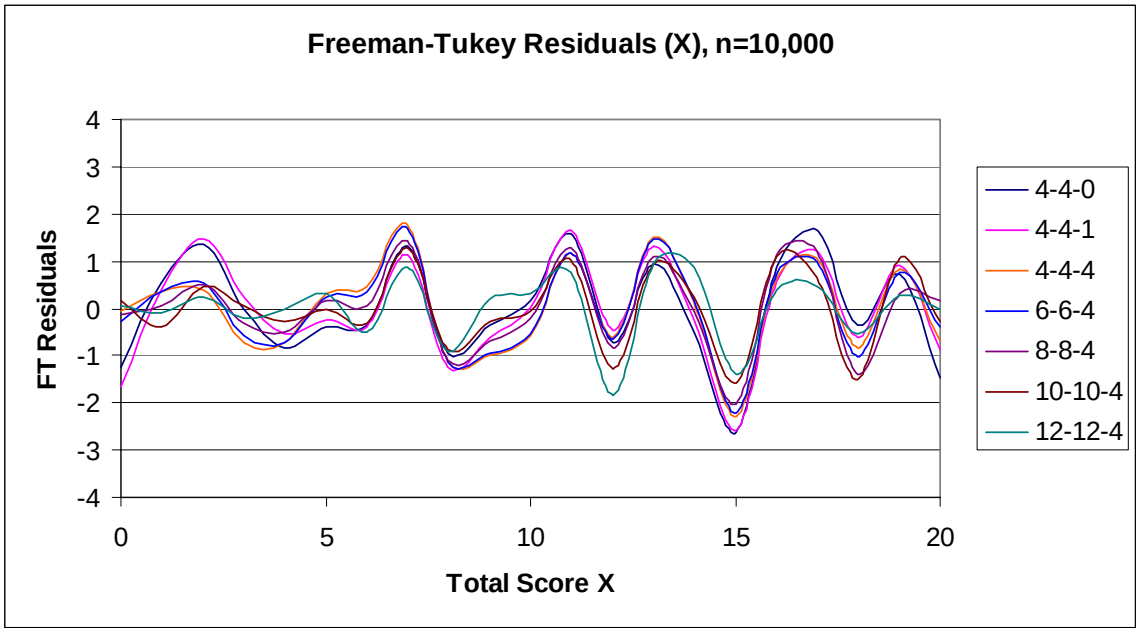


Figure A.8
20 Items per form, n=10,000



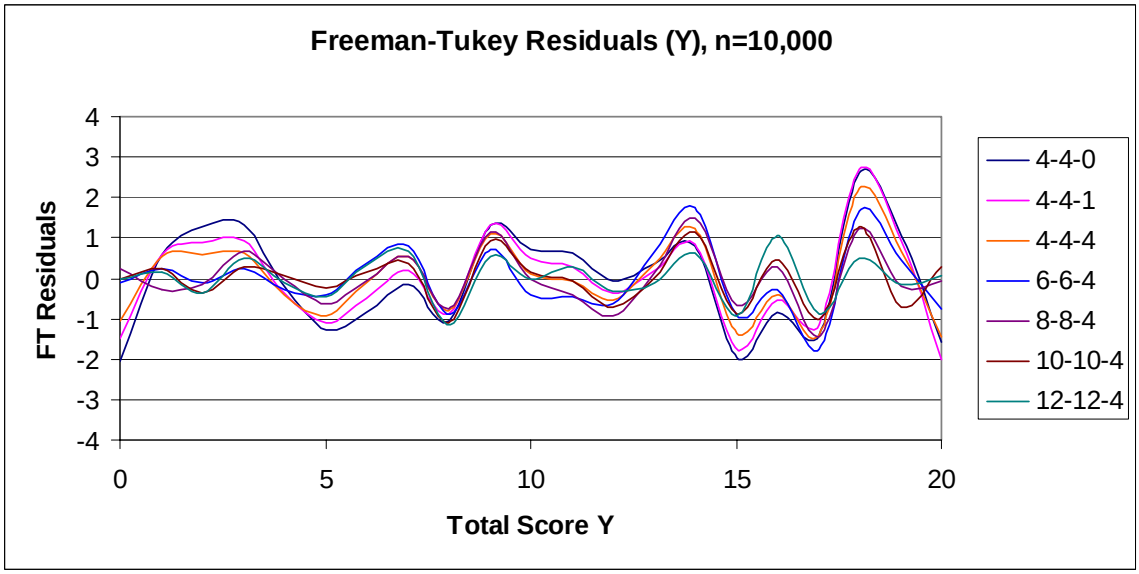
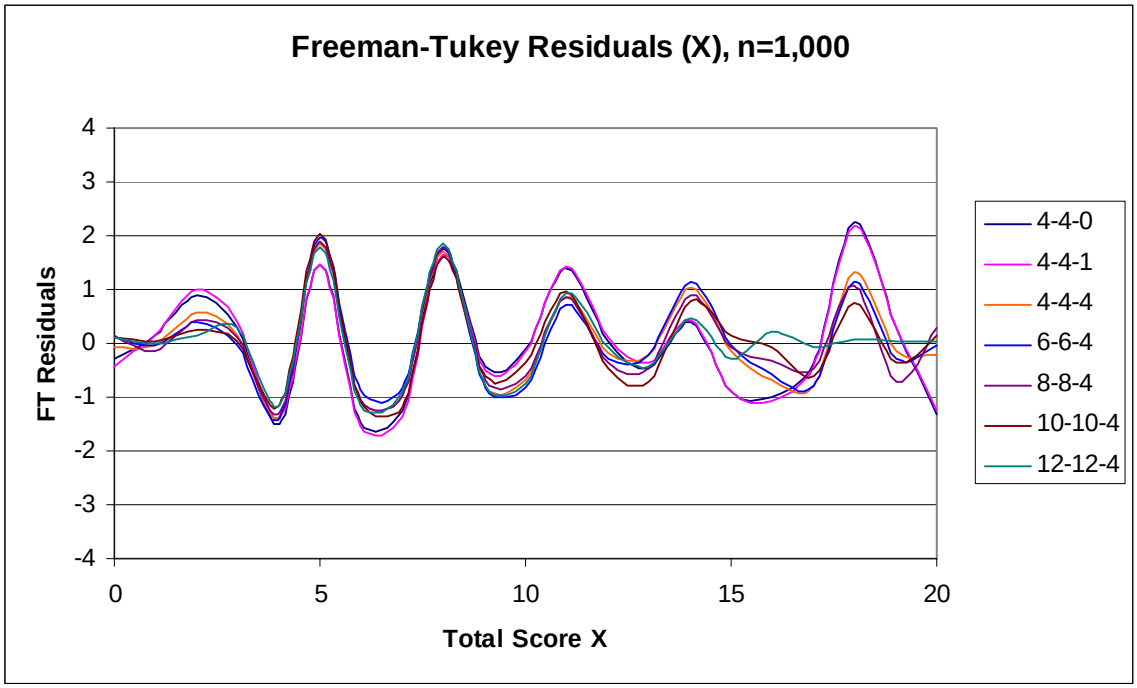
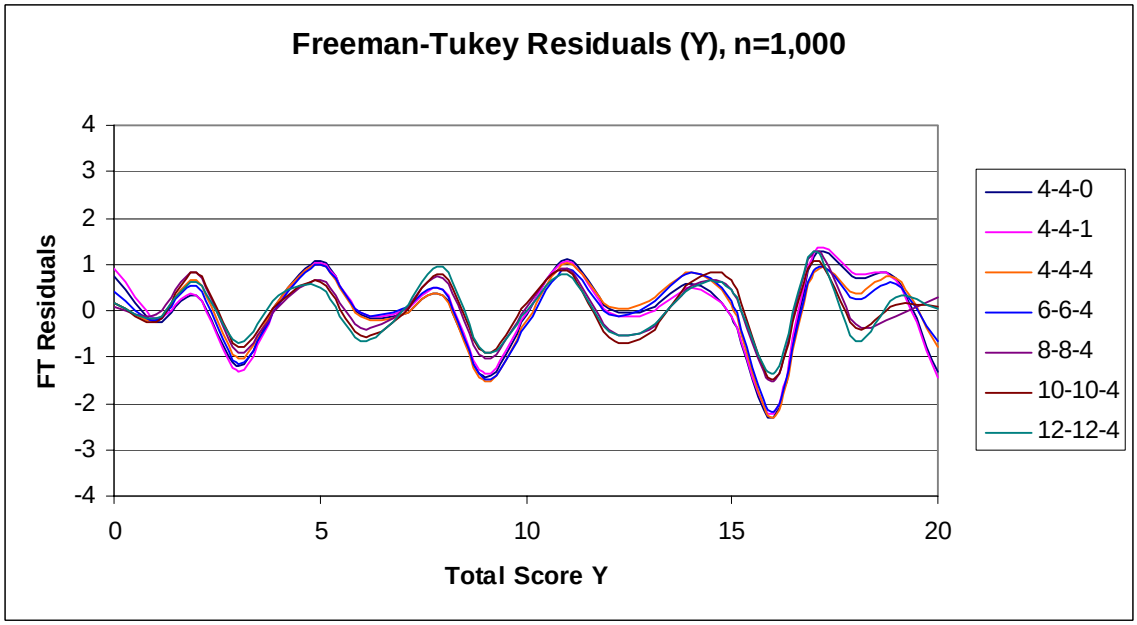


Figure A.9
20 Items per form, n=1000





APPENDIX B: Equating Functions

Figure B.1
100 Items per form

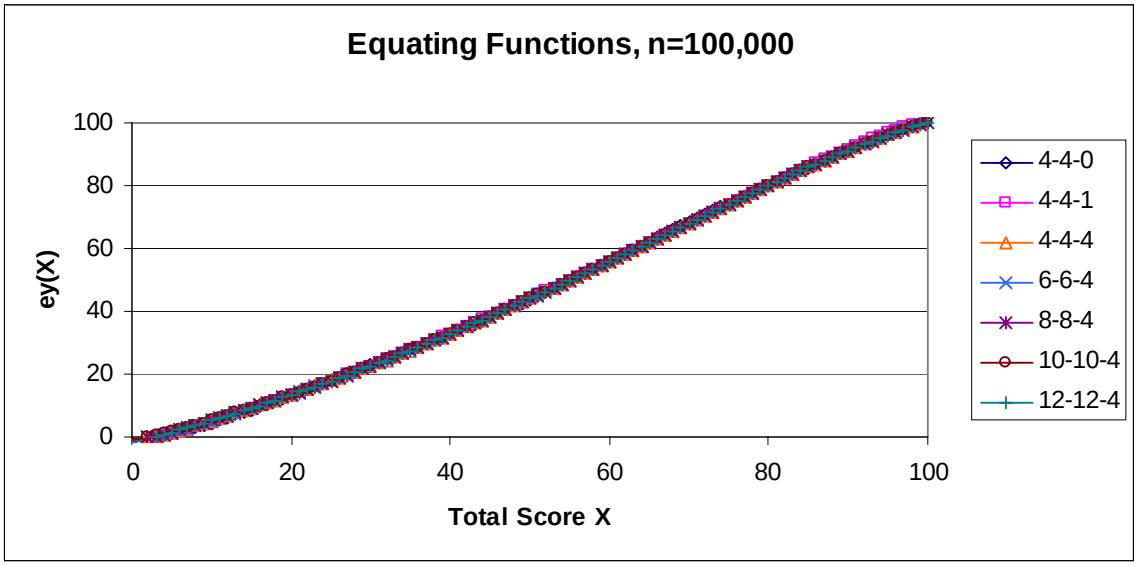


Figure B.2
100 Items per form

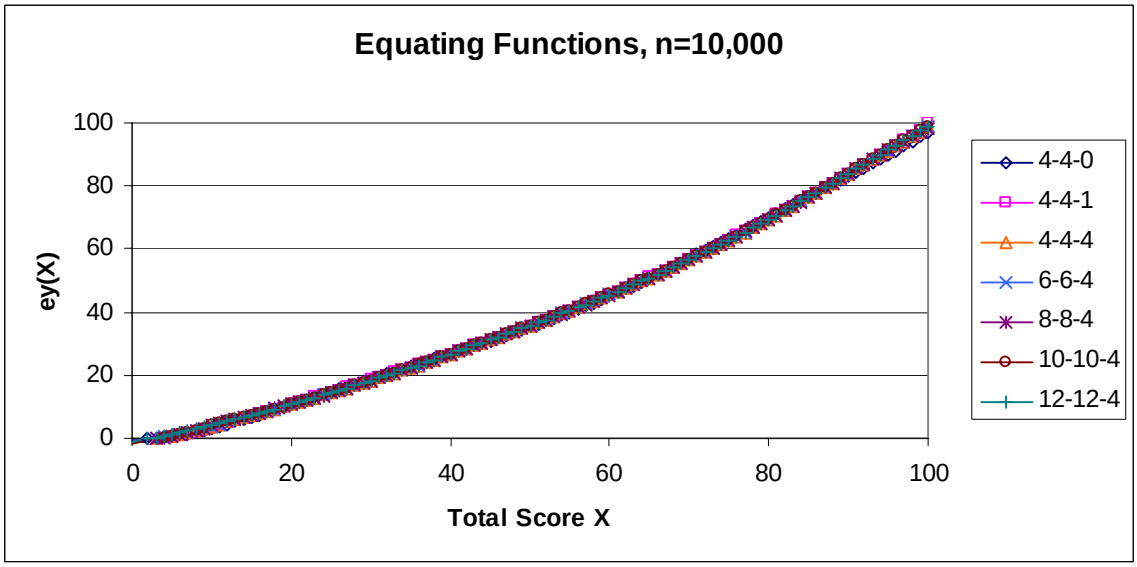


Figure B.3
100 Items per form

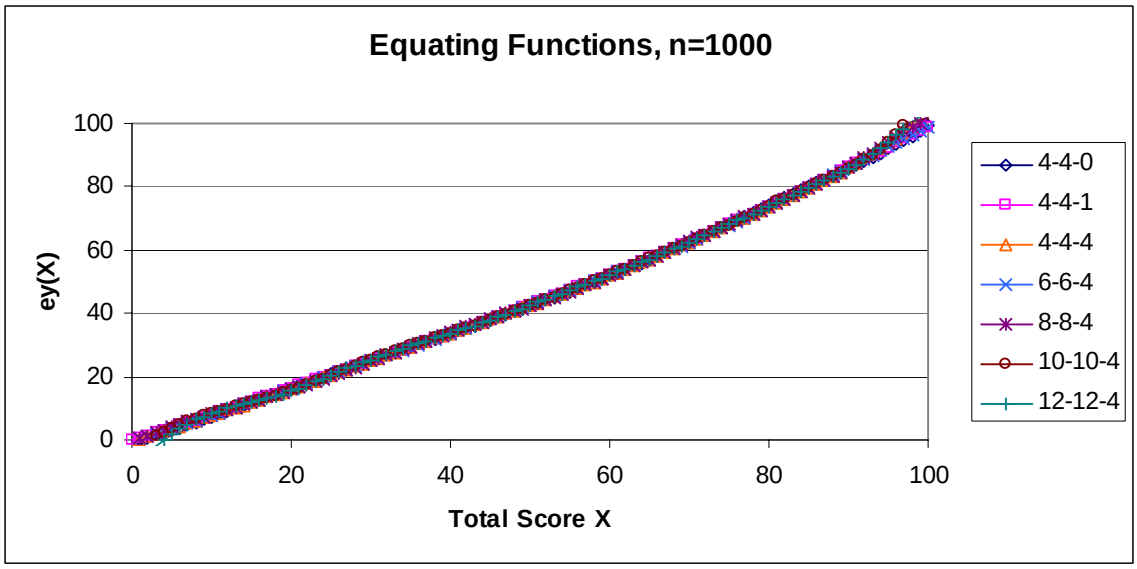


Figure B.4
60 Items per form

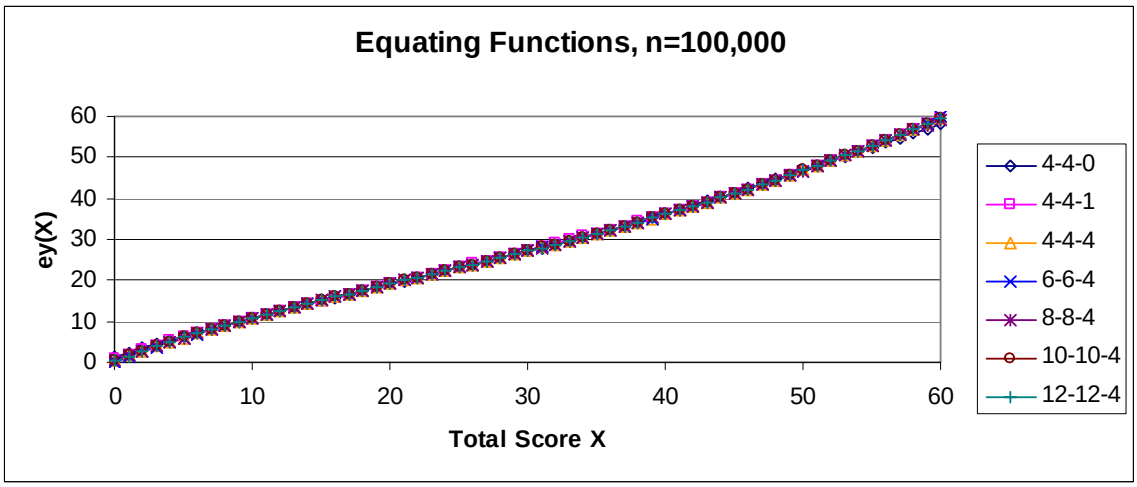


Figure B.5
60 Items per form

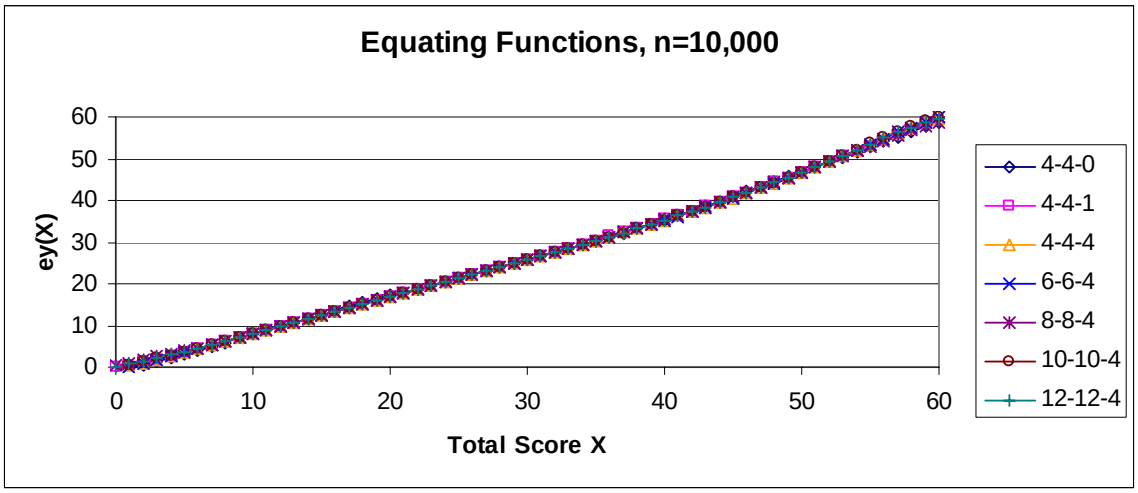


Figure B.6
60 Items per form

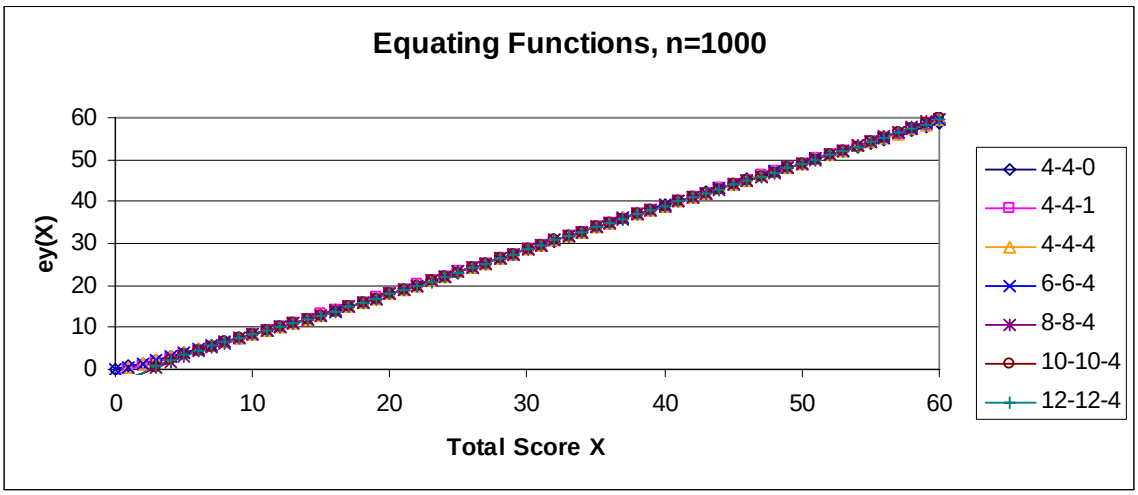


Figure B.7
20 Items per form

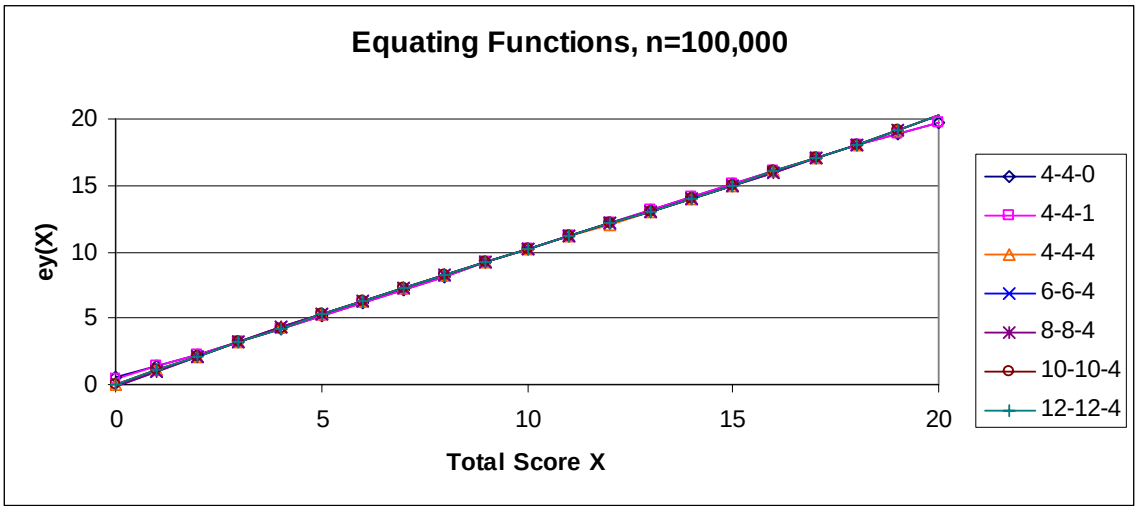


Figure B.8
20 Items per form

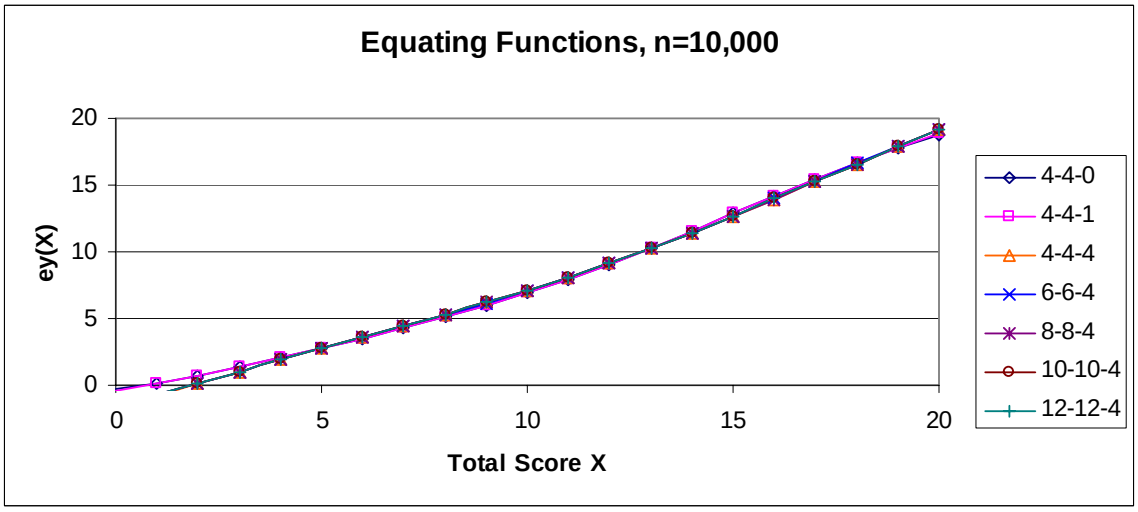
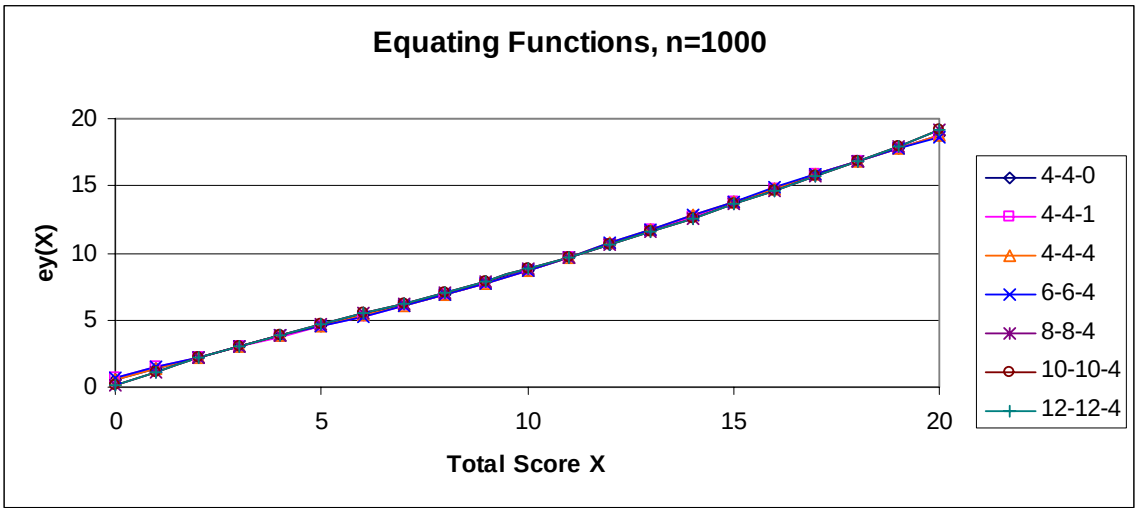


Figure B.9
20 Items per form



APPENDIX C: Equating Function Differences

Figure C.1
100 Items per form, n=100,000

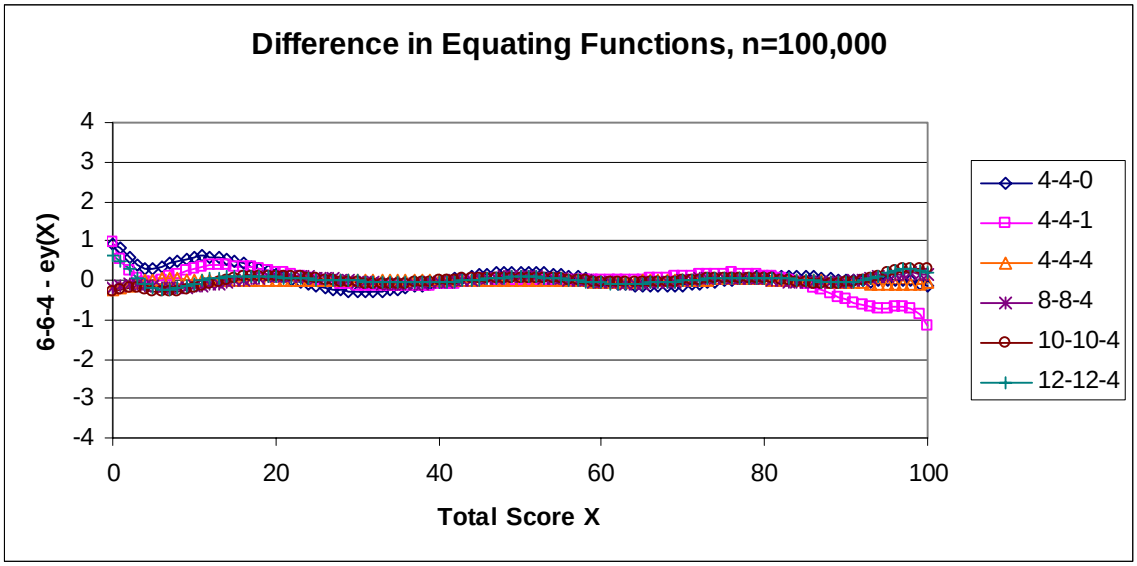


Figure C.2
100 Items per form, n=10,000

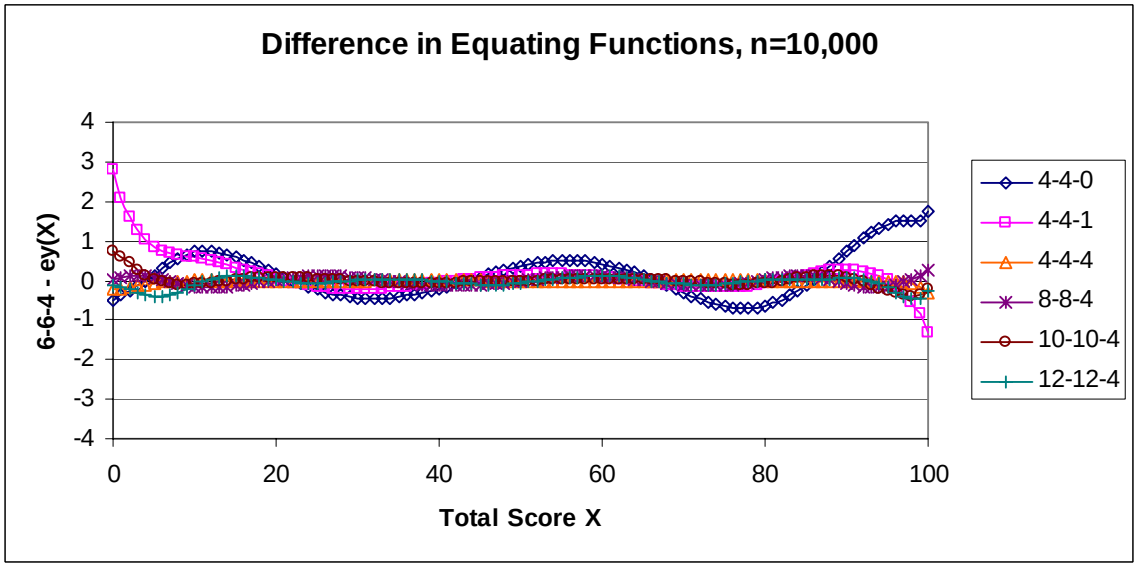


Figure C.3
100 Items per form, n=1000

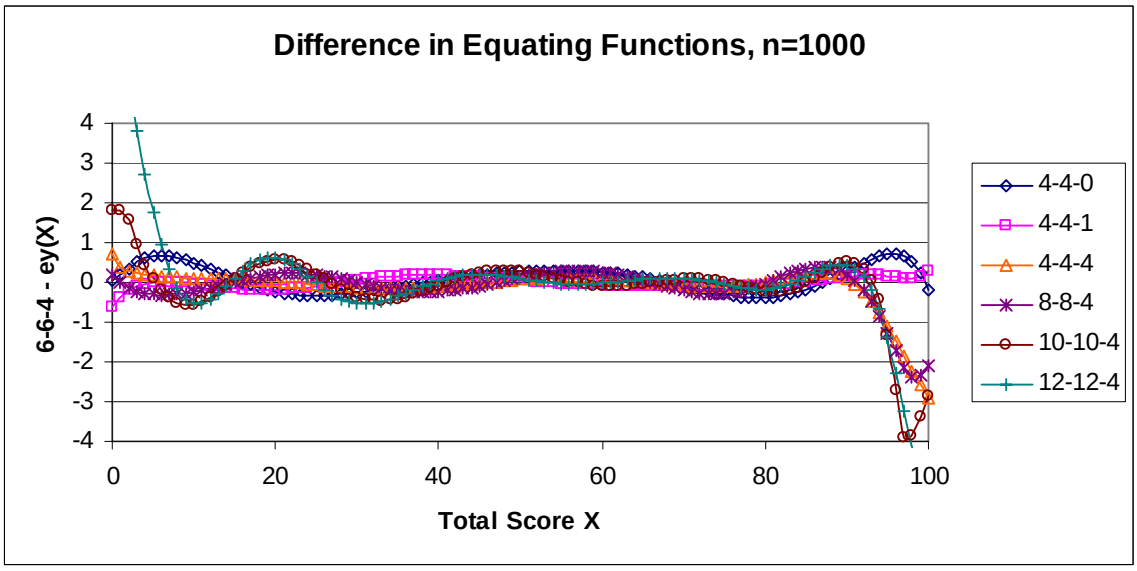


Figure C.4
60 Items per form, n=100,000

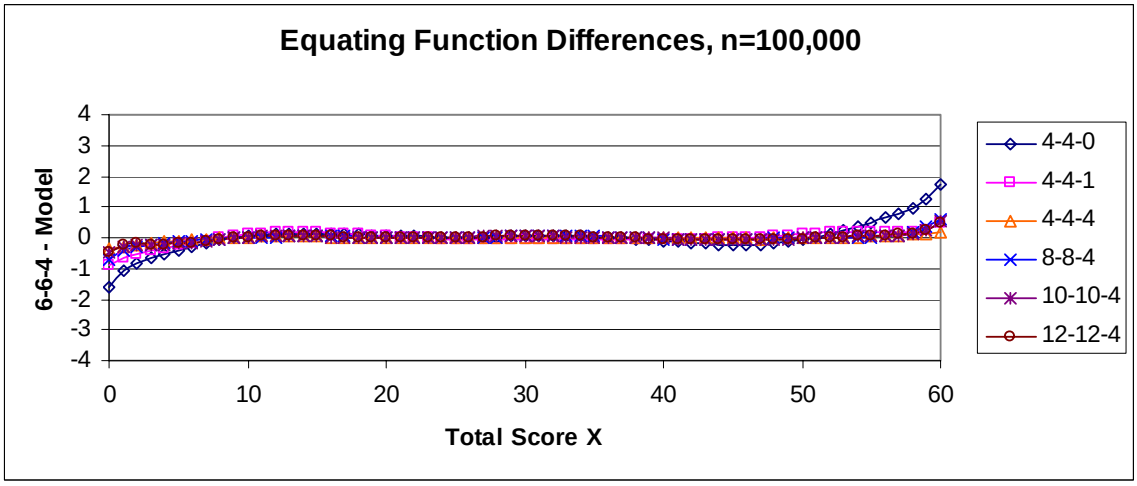


Figure C.5
60 Items per form, n=10,000

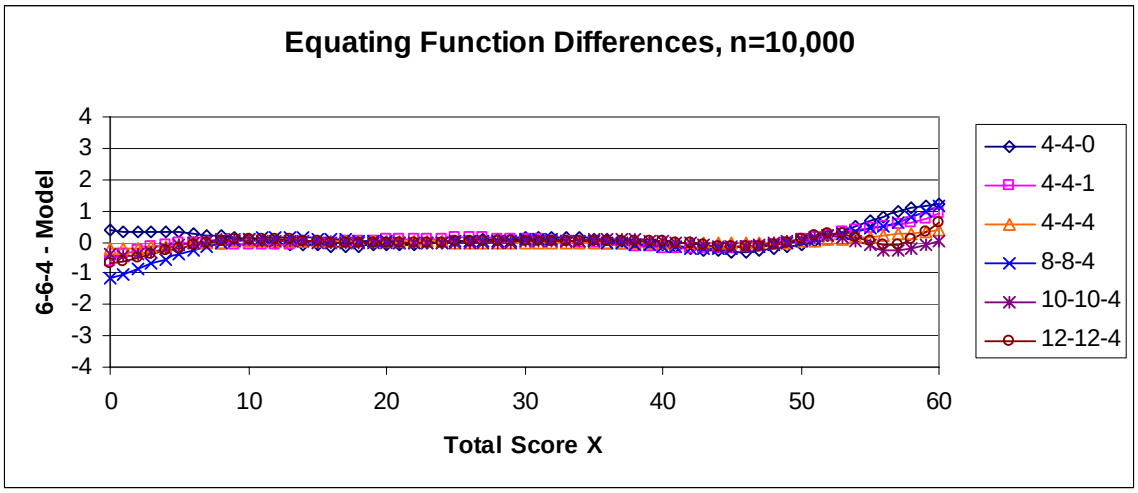


Figure C.6
60 Items per form, n=1000

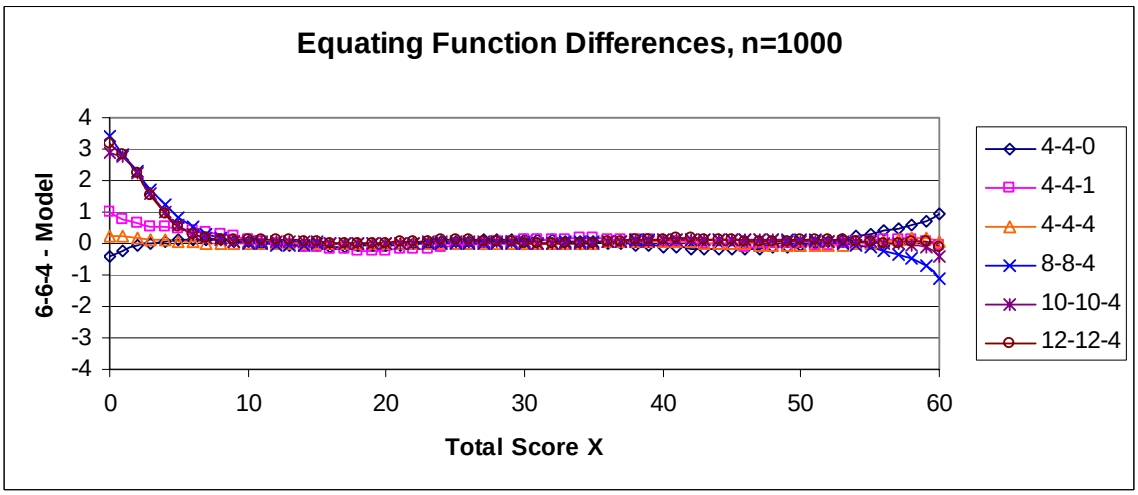


Figure C.7
20 Items per form, n=100,000

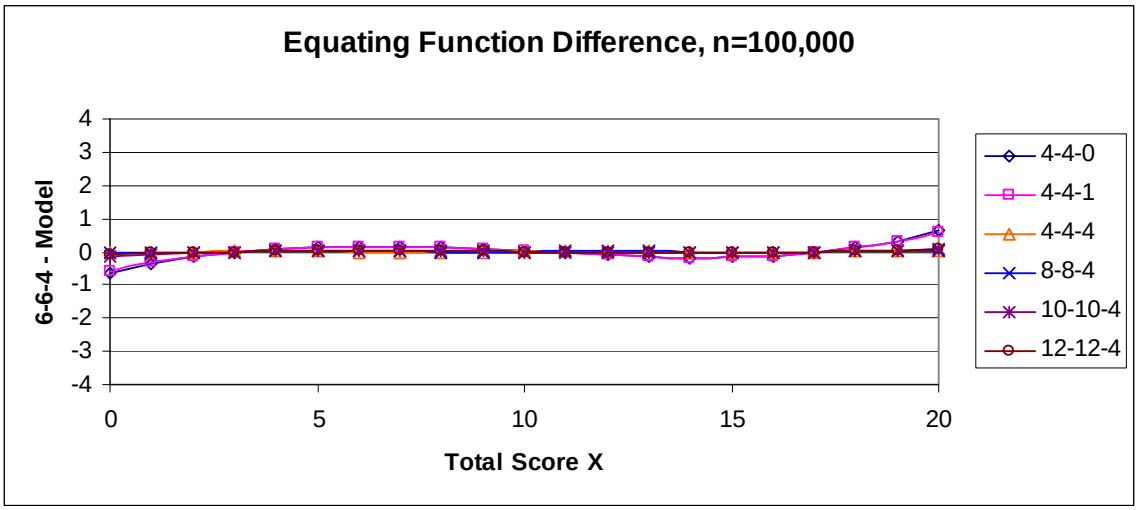


Figure C.8
20 Items per form, n=10,000

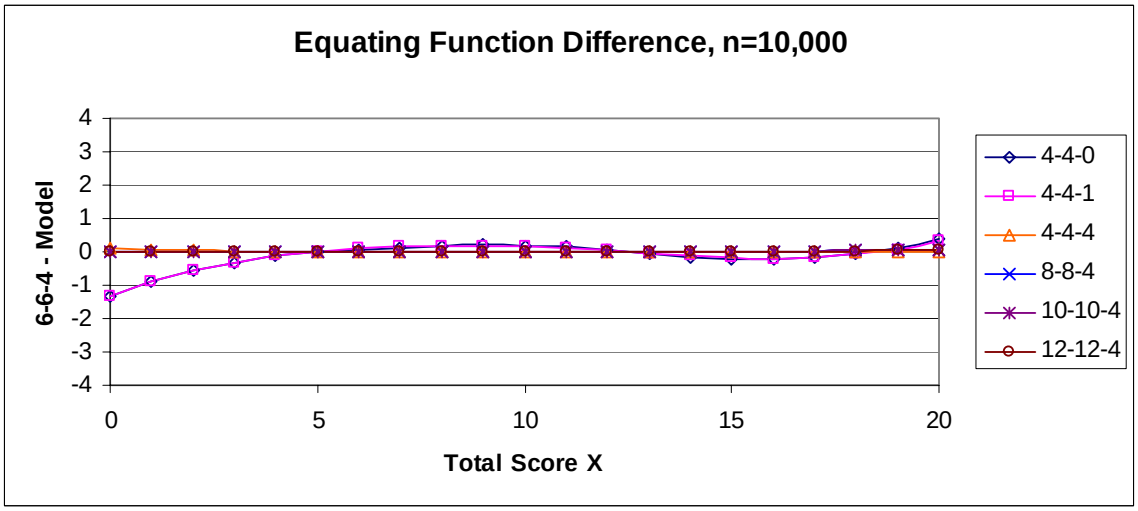
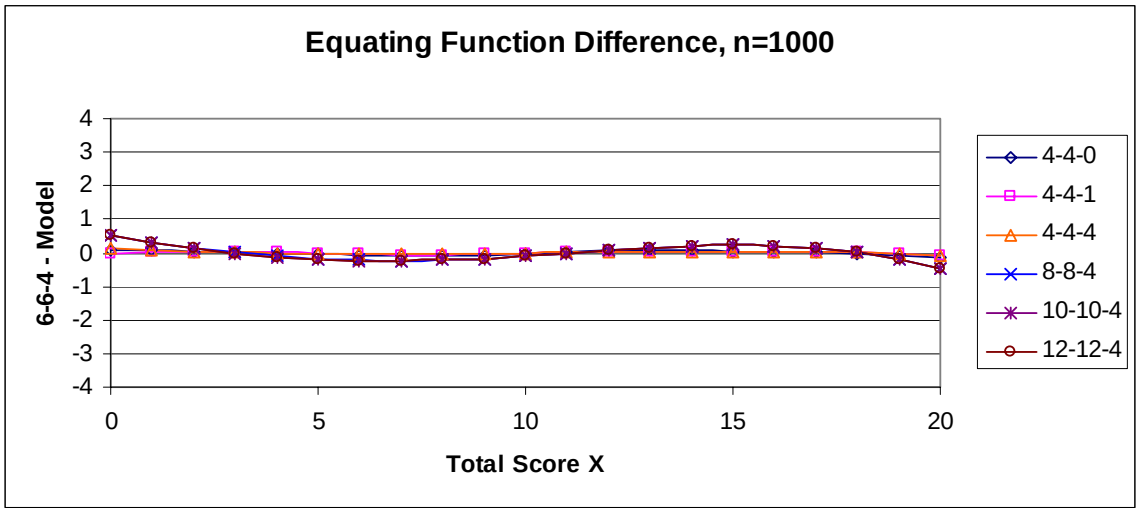


Figure C.9
20 Items per form, n=1000



APPENDIX D: Standard Errors of Equating

Figure D.1: 100 Items per form, n=100,000

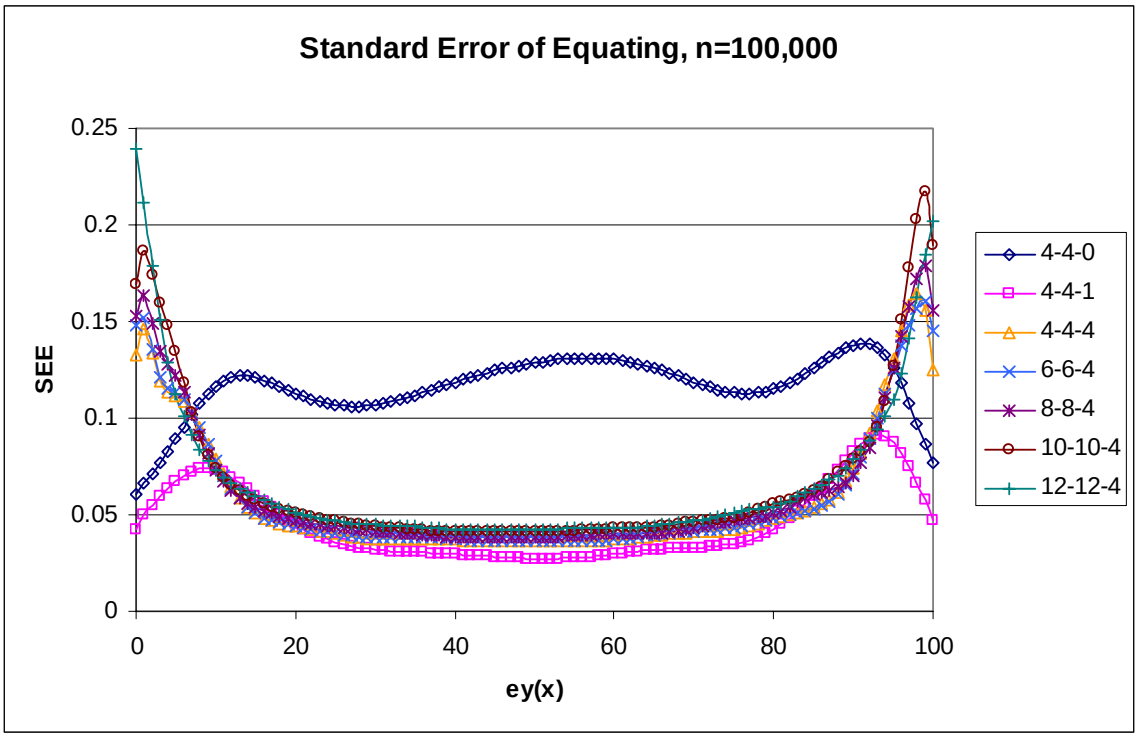


Figure D.2: 100 Items per form, n=10,000

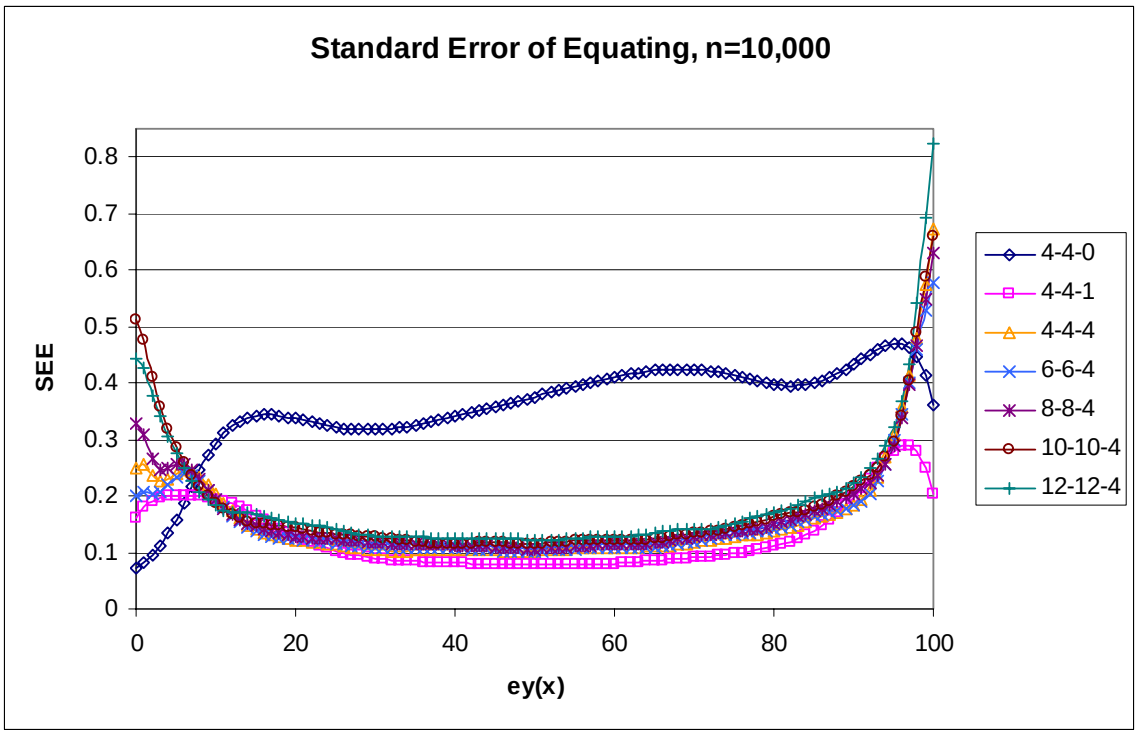


Figure D.3: 100 Items per form, n=1000

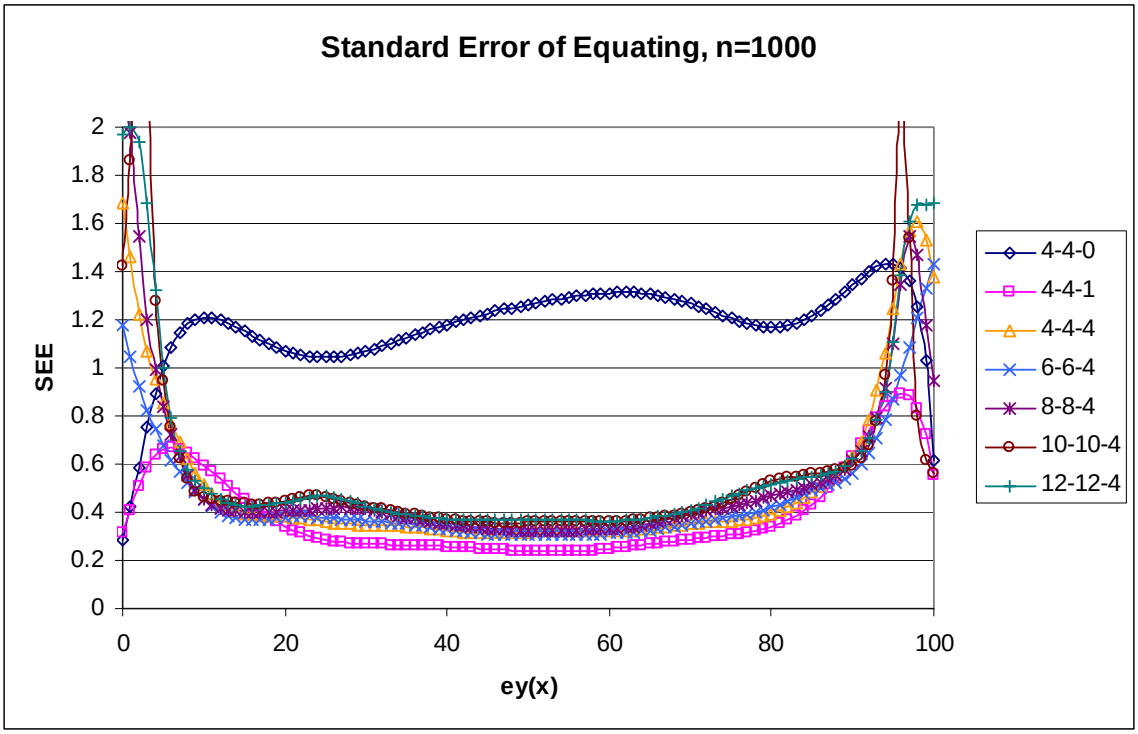


Figure D.4: 60 Items per form, n=100,000

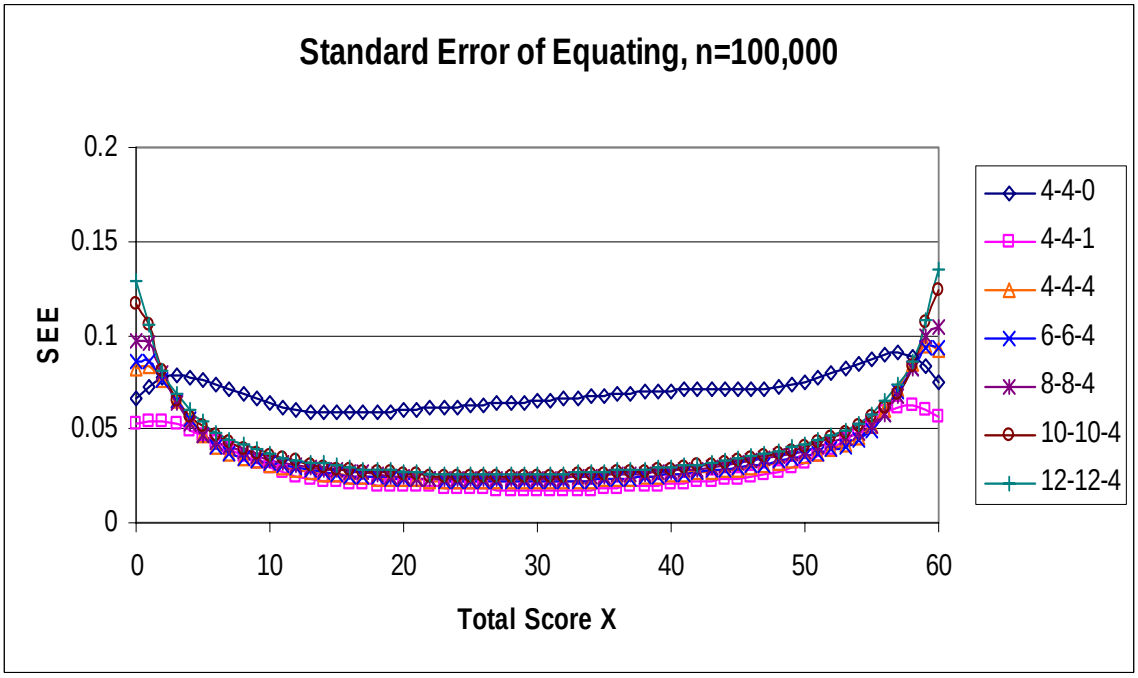


Figure D.5: 60 Items per form, n=10,000

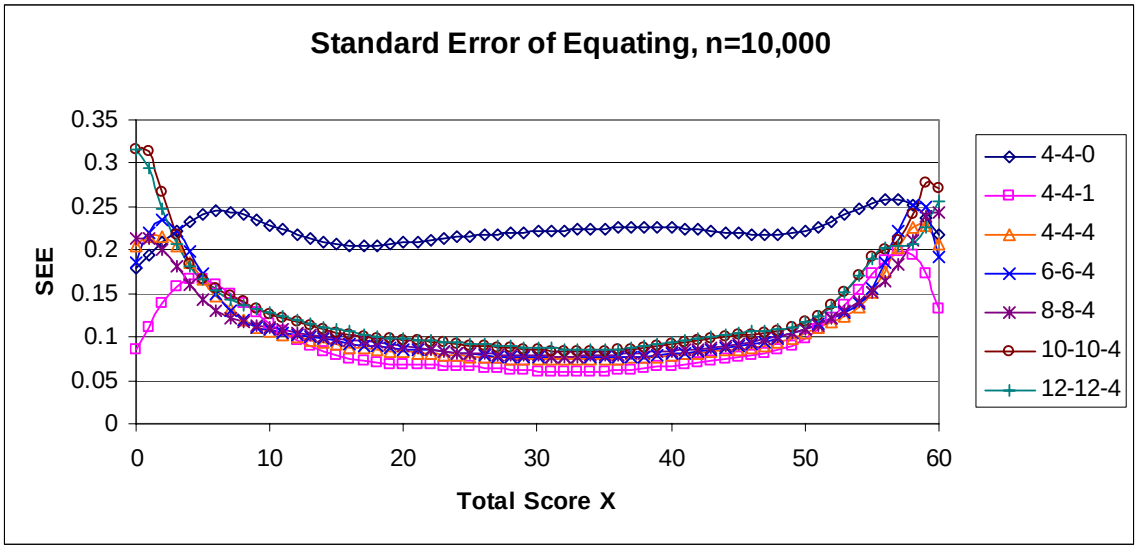


Figure D.6: 60 Items per form, n=1000

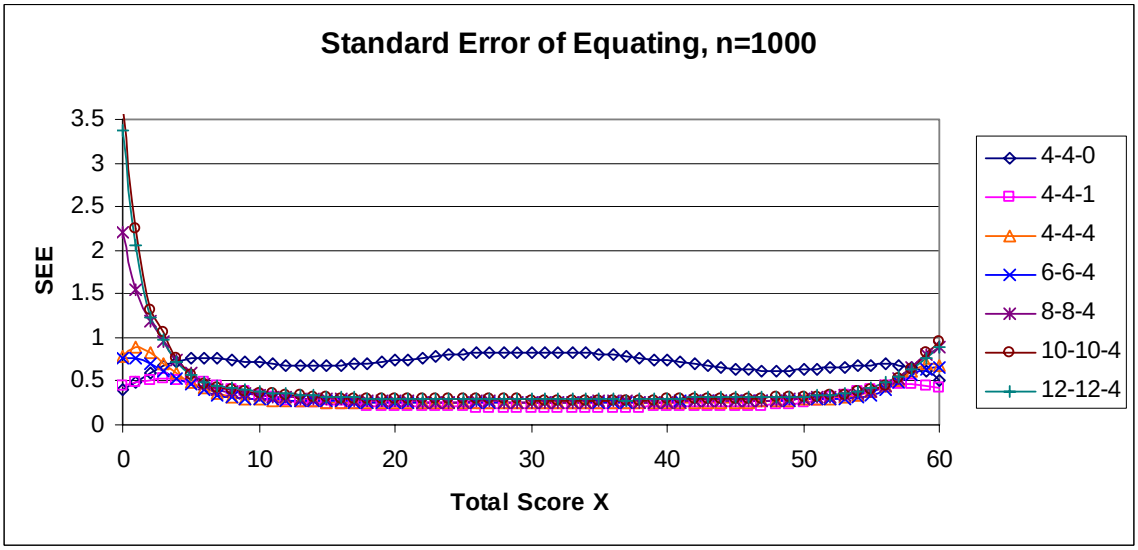


Figure D.7: 20 Items per form, n=100,000

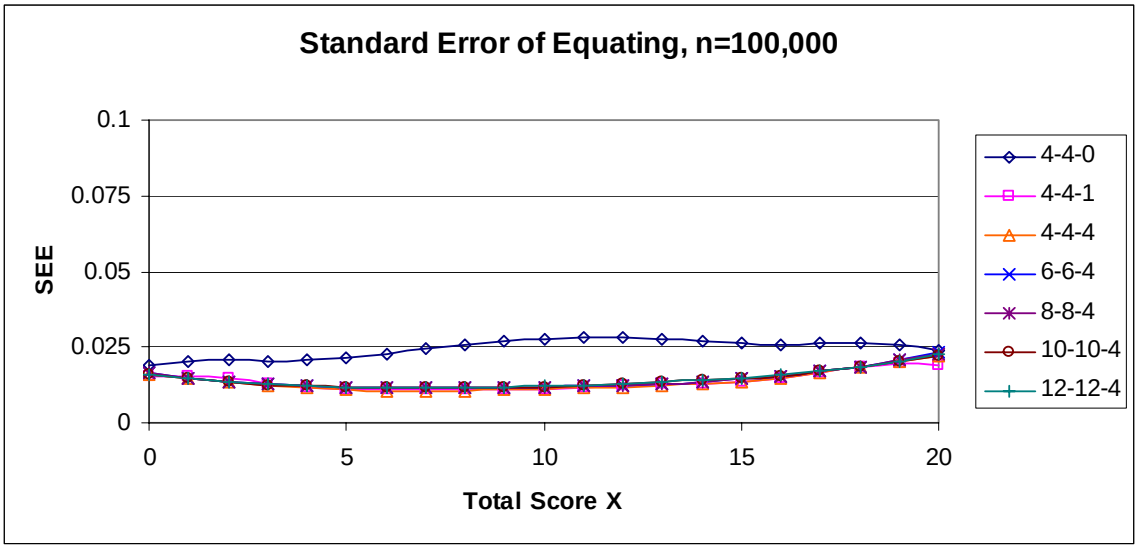


Figure D.8: 20 Items per form, n=10,000

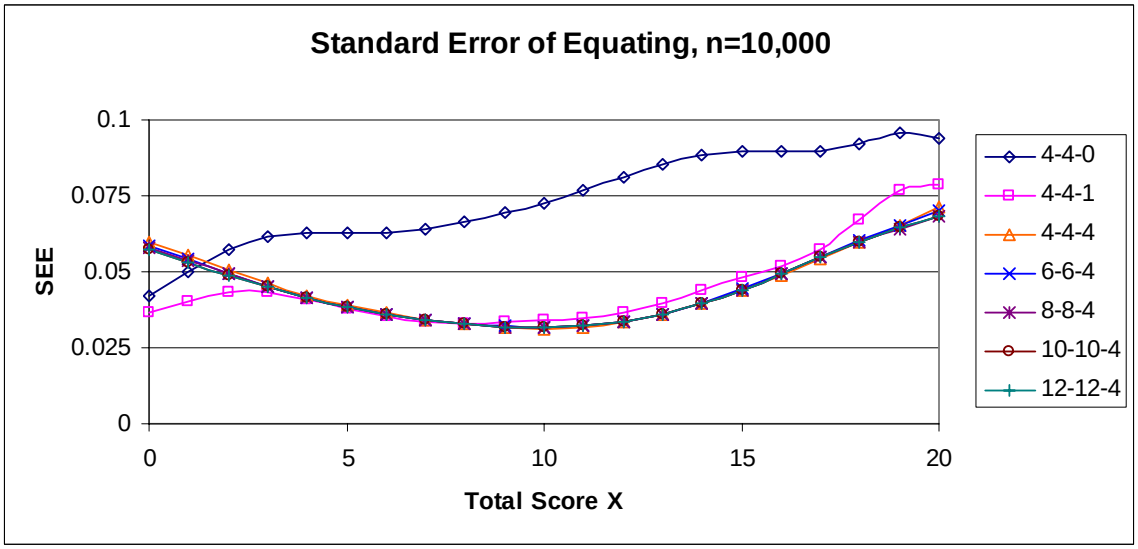
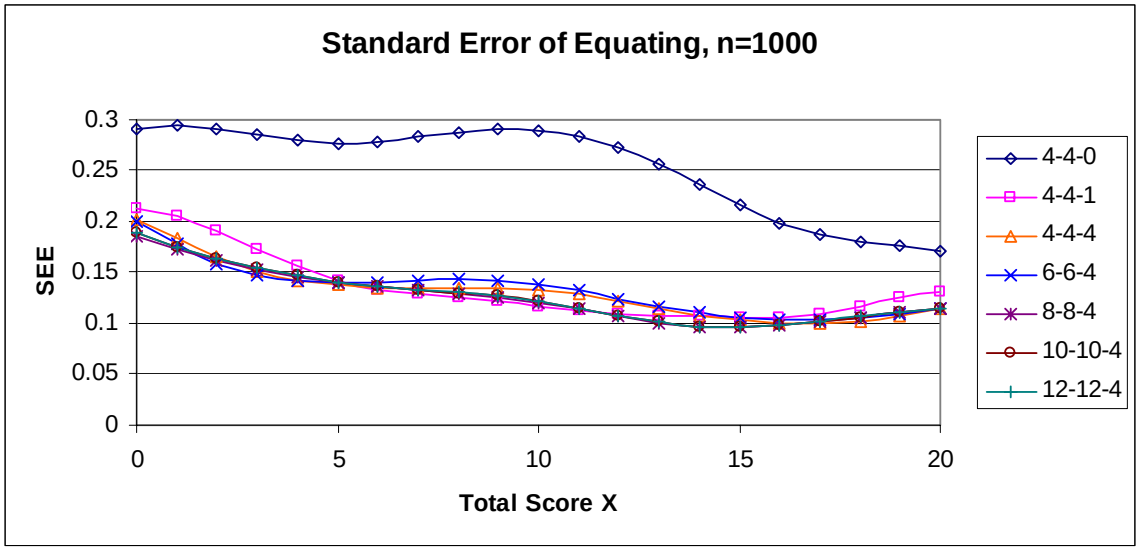


Figure D.9: 20 Items per form, n=1000



Appendix E: Equating Functions

Figure E.1: 100 Items per form, 50% Anchor Length, 1000 Sample Size, No Ability Difference

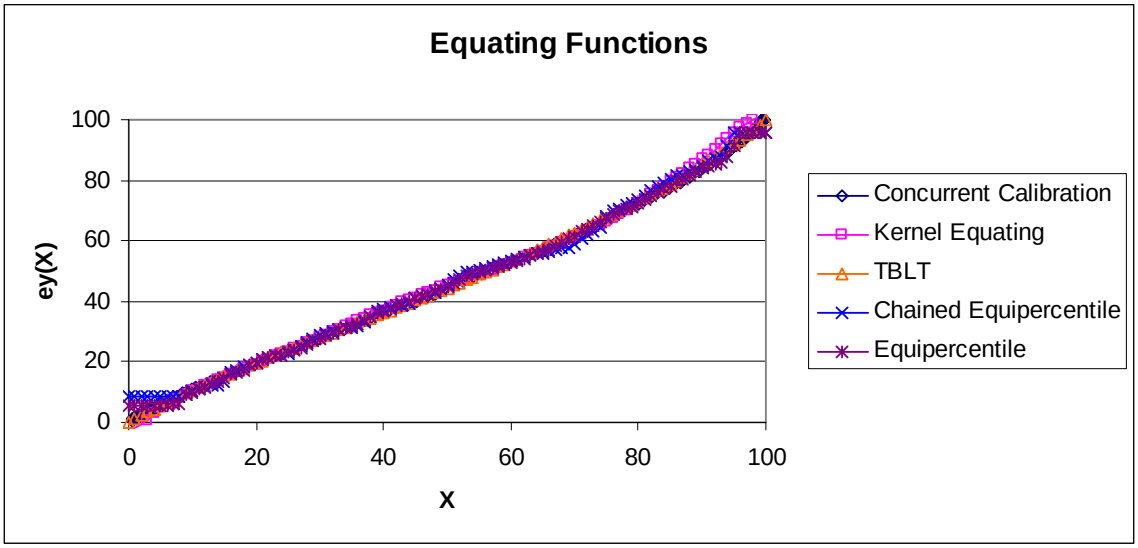


Figure E.2: 100 Items per form, 50% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

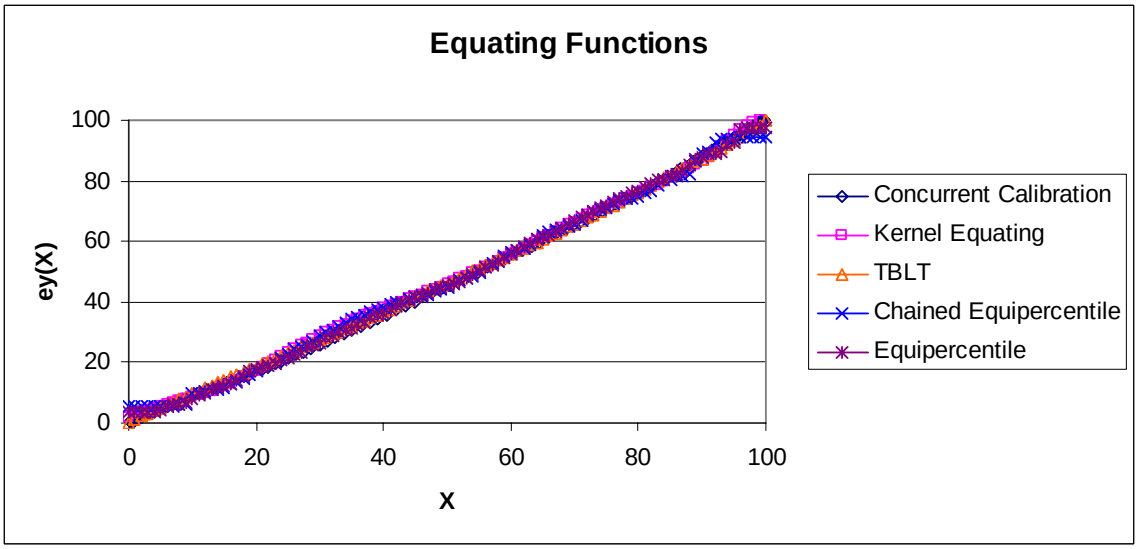


Figure E.3: 100 Items per form, 50% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

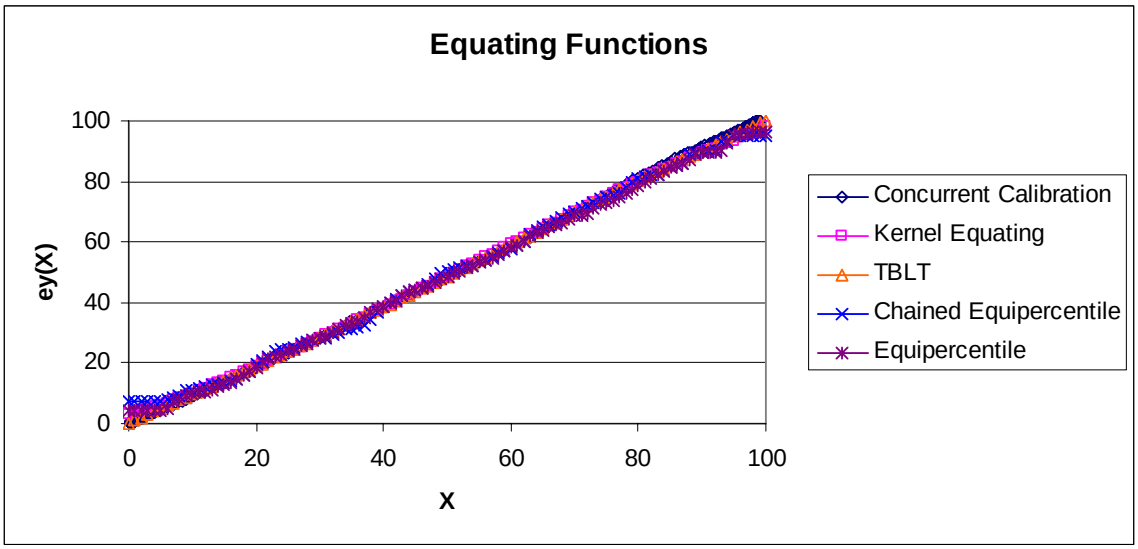


Figure E.4: 100 Items per form, 50% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

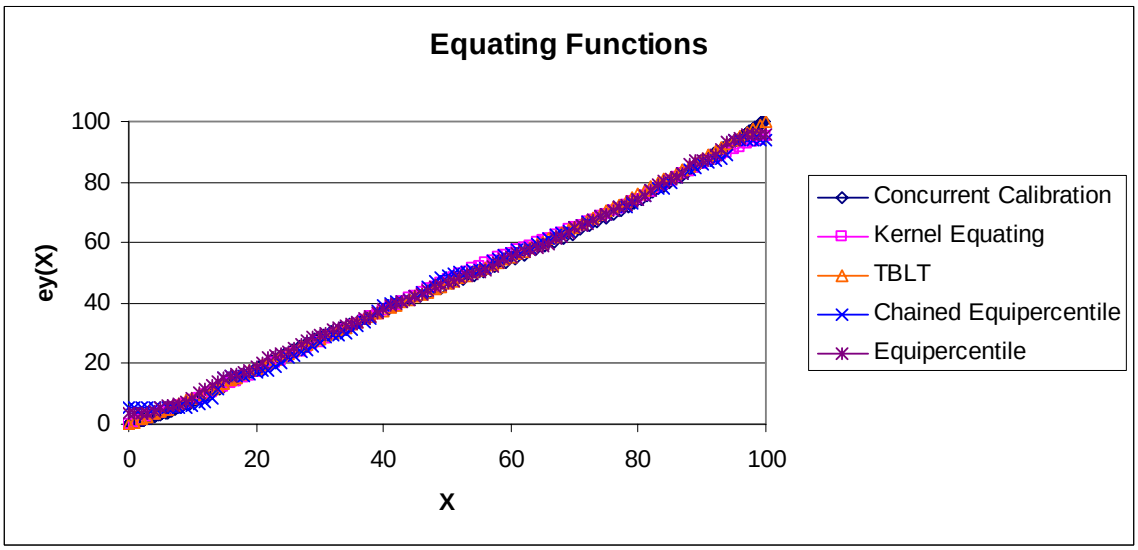


Figure E.5: 100 Items per form, 50% Anchor Length, 10,000 Sample Size, No Ability Difference

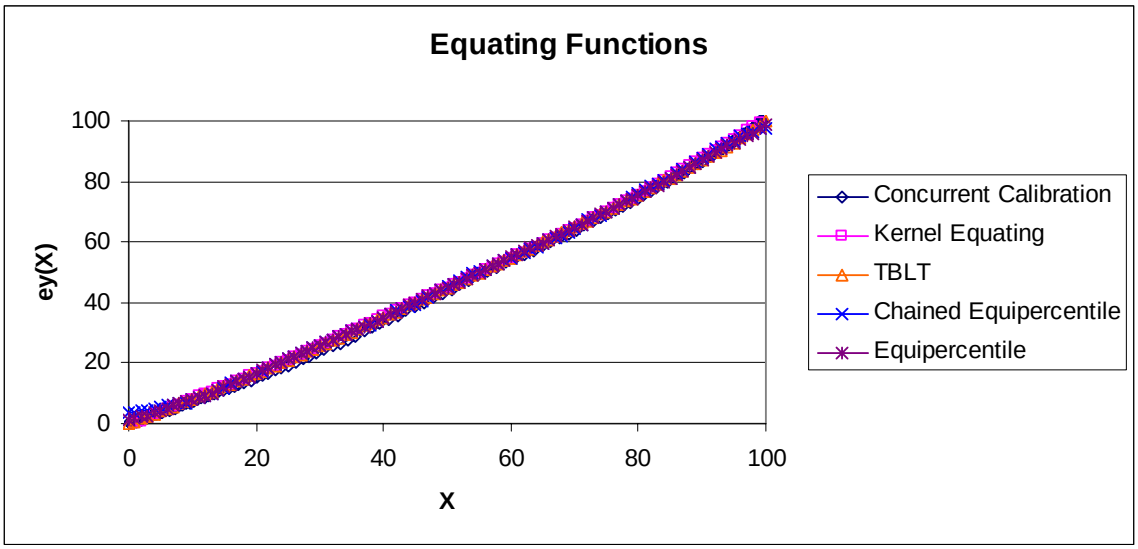


Figure E.6: 100 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

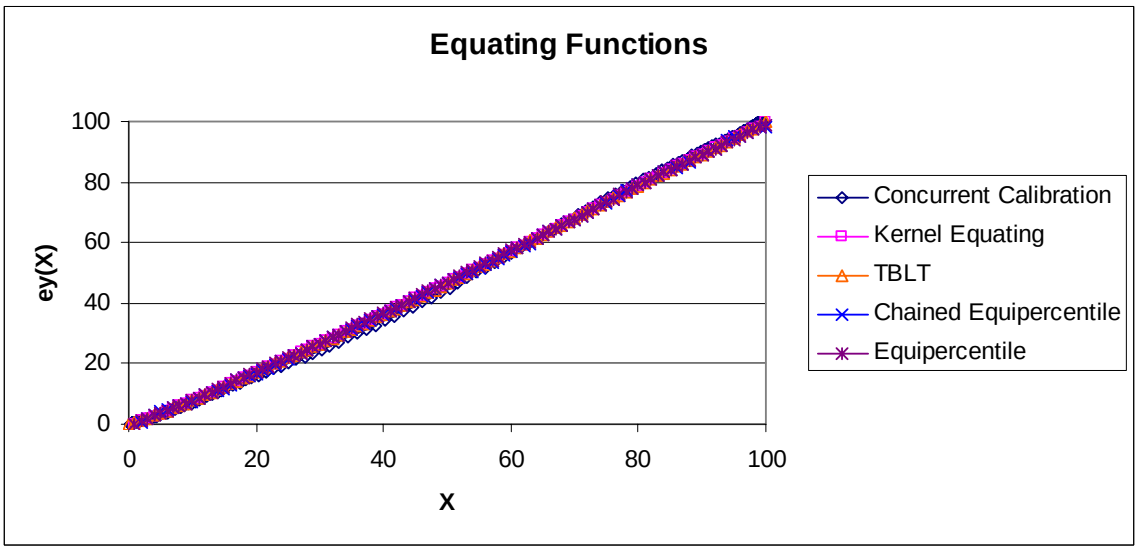


Figure E.7: 100 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

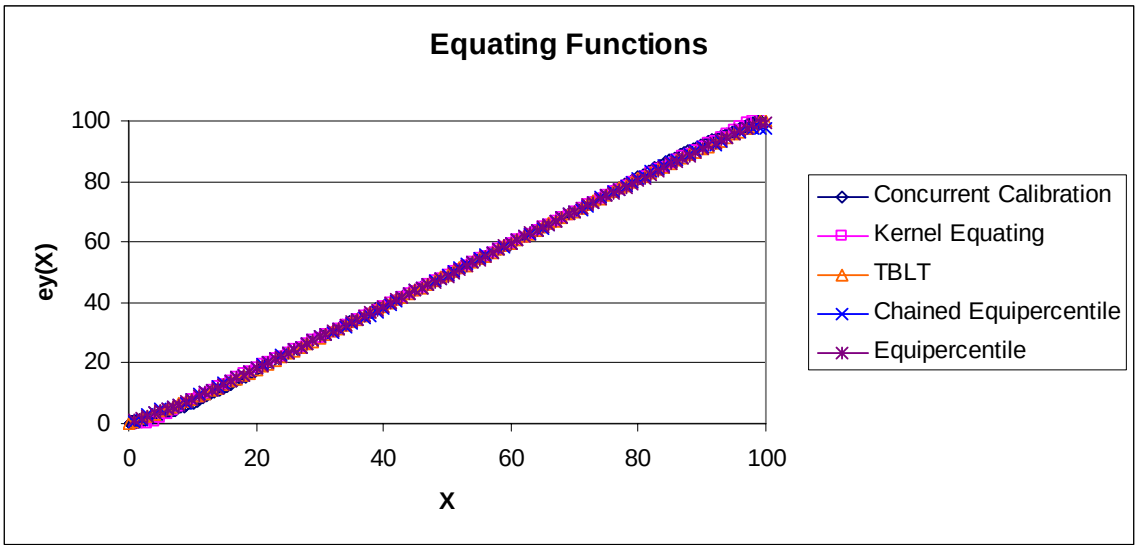


Figure E.8: 100 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

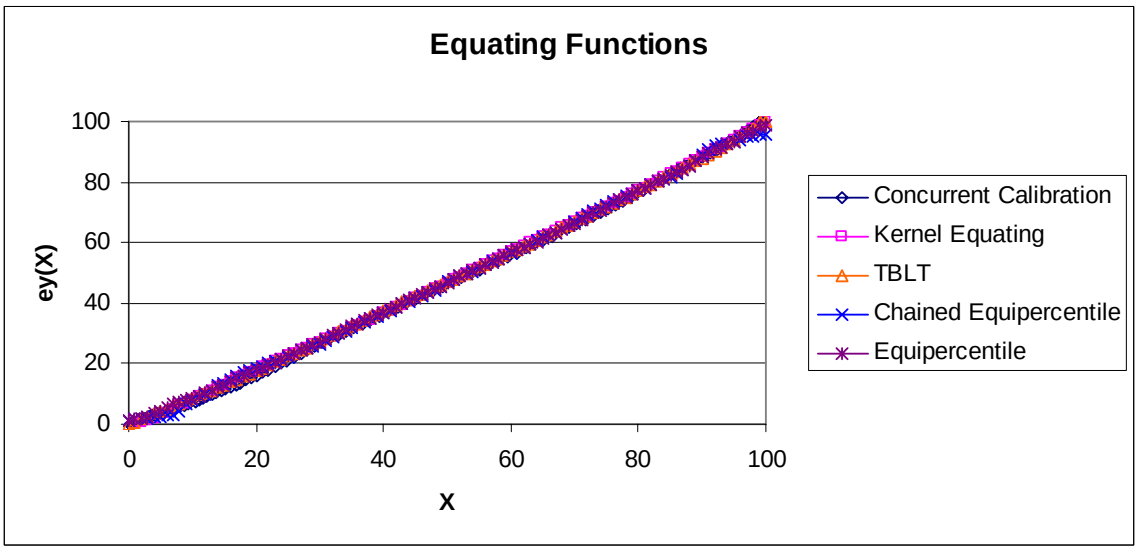


Figure E.9: 100 Items per form, 50% Anchor Length, 100,000 Sample Size, No Ability Difference

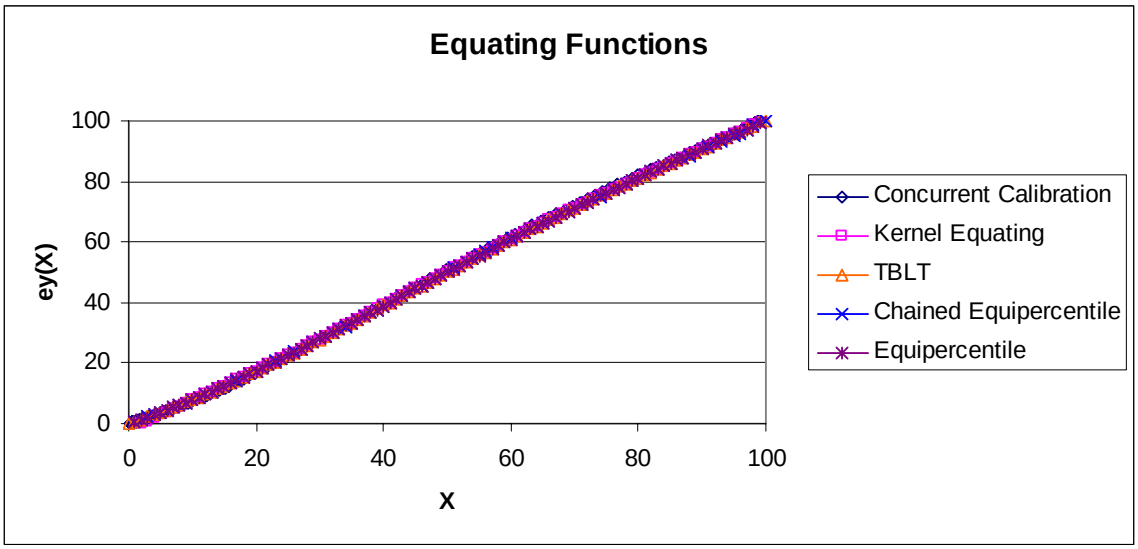


Figure E.10: 100 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

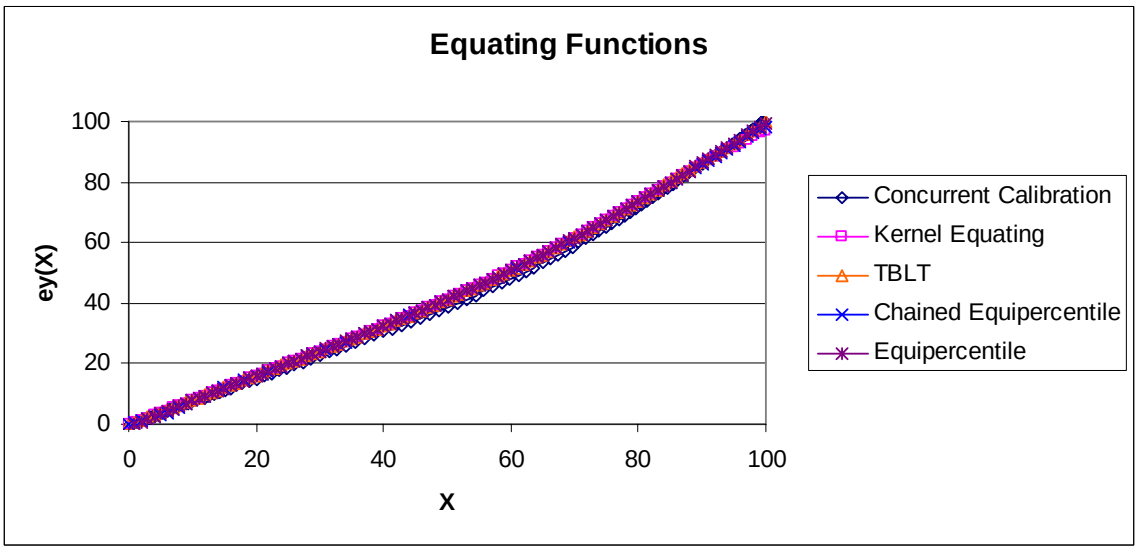


Figure E.11: 100 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

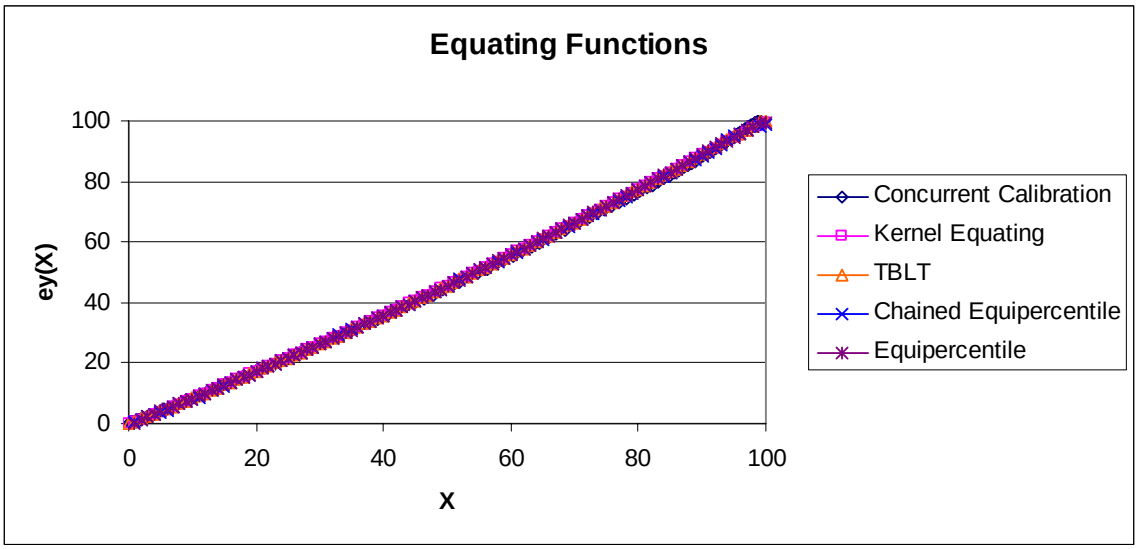


Figure E.12: 100 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

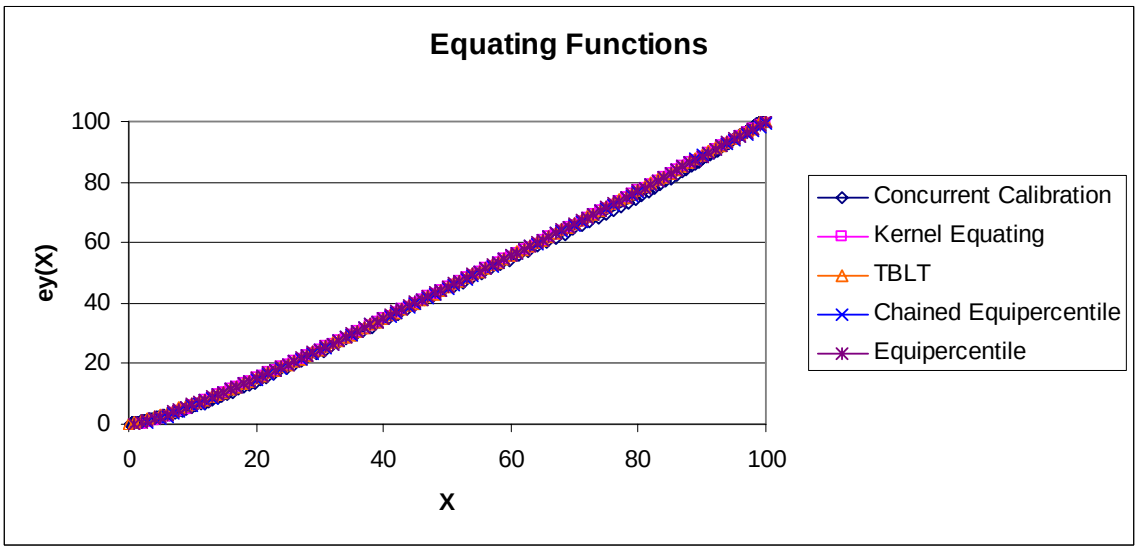


Figure E.13: 100 Items per form, 35% Anchor Length, 1000 Sample Size, No Ability Difference

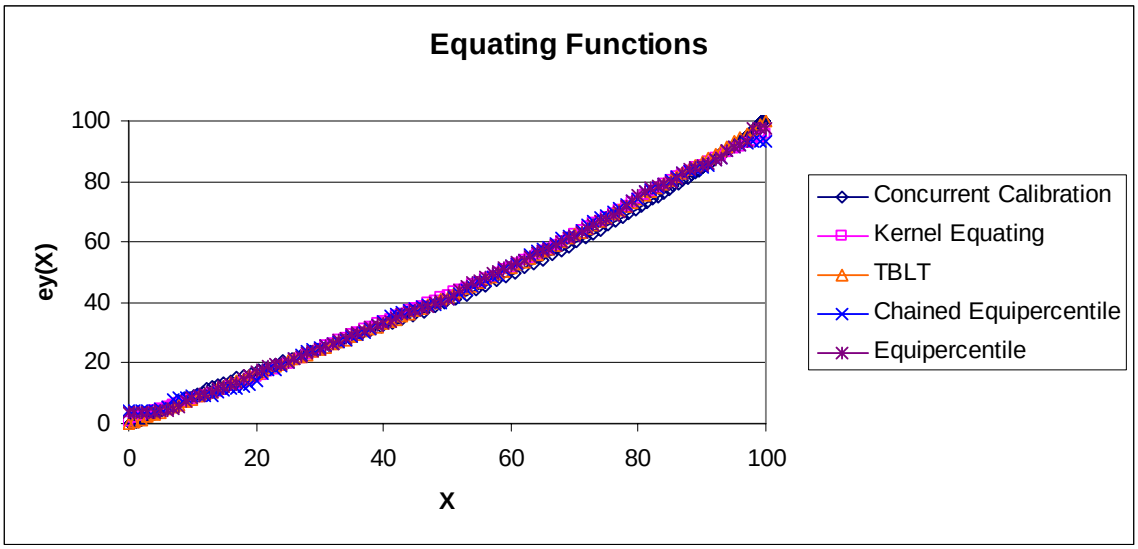


Figure E.14: 100 Items per form, 35% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

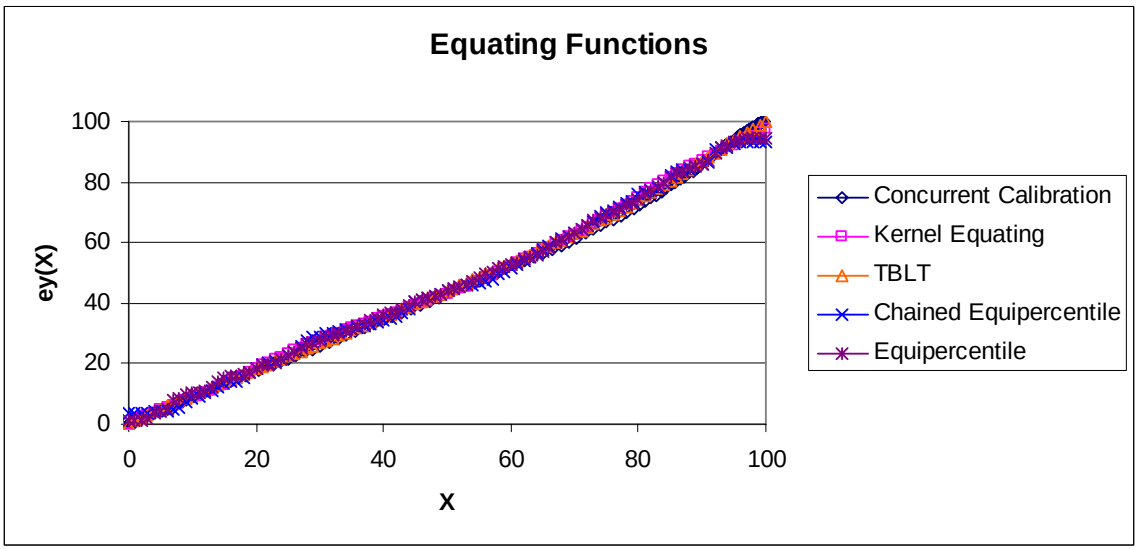


Figure E.15: 100 Items per form, 35% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

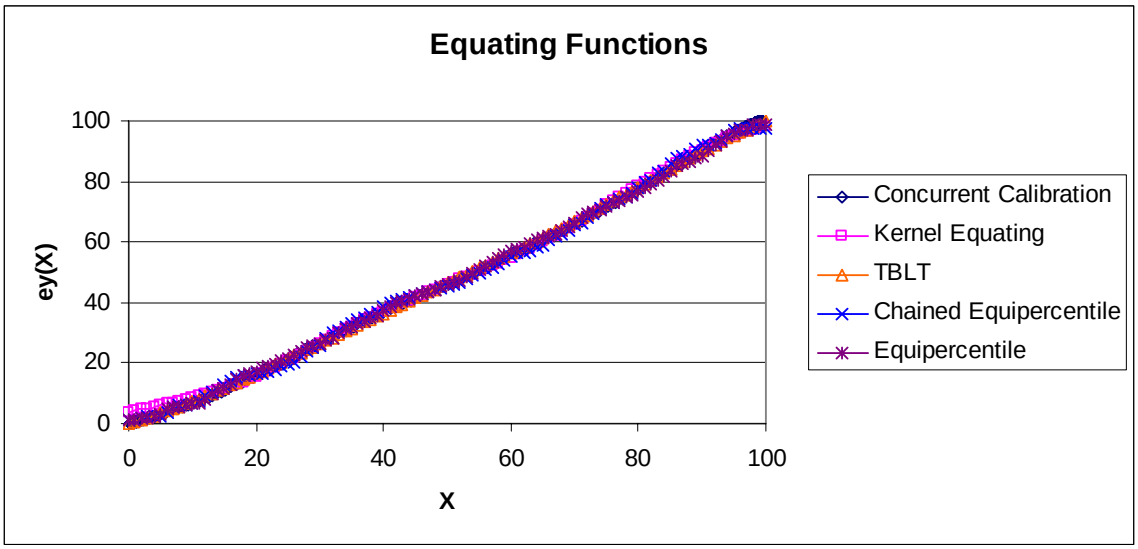


Figure E.16: 100 Items per form, 35% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

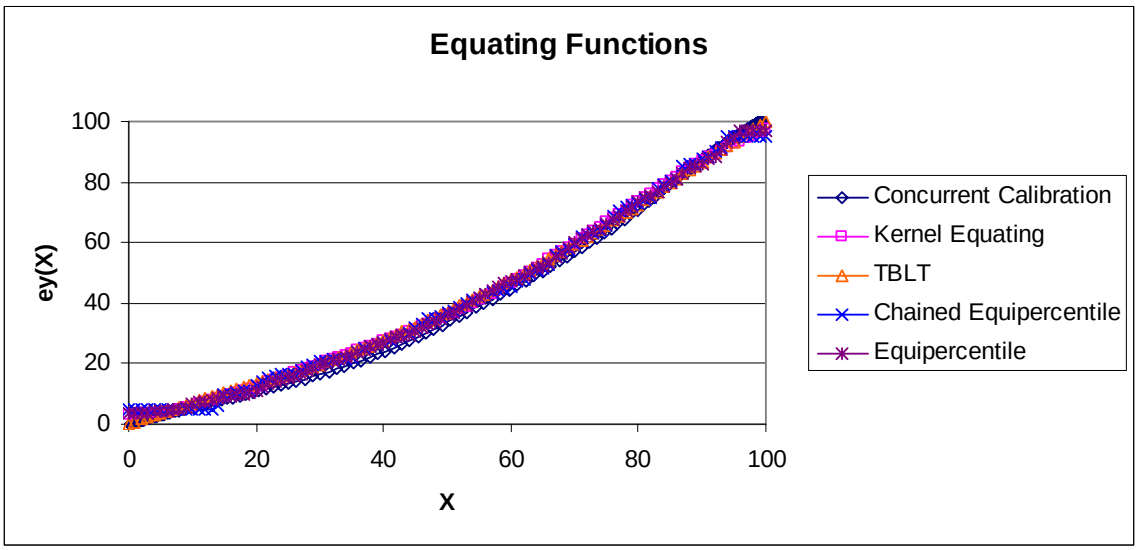


Figure E.17: 100 Items per form, 35% Anchor Length, 10,000 Sample Size, No Ability Difference

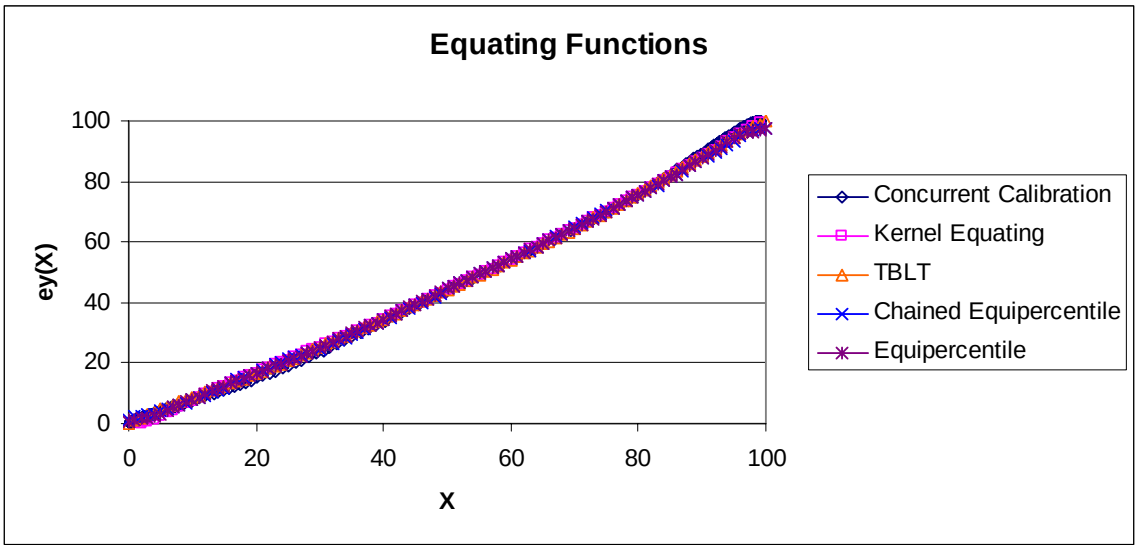


Figure E.18: 100 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

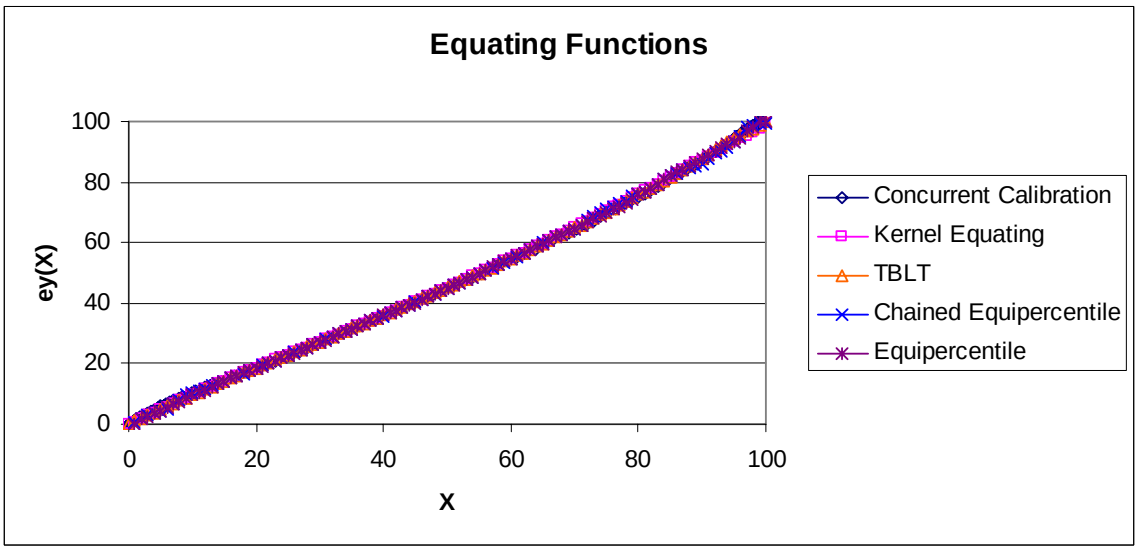


Figure E.19: 100 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

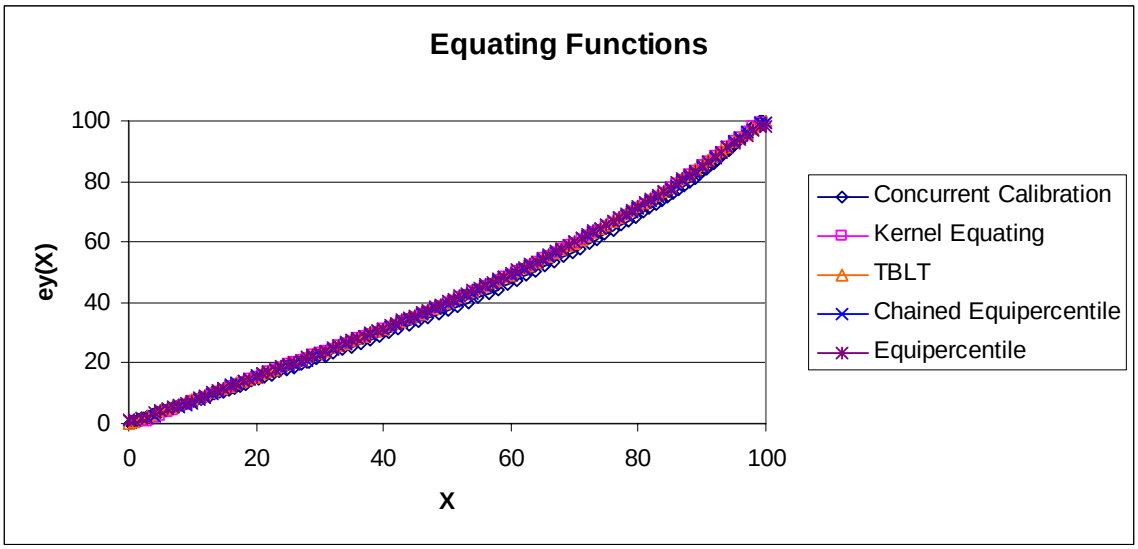


Figure E.20: 100 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

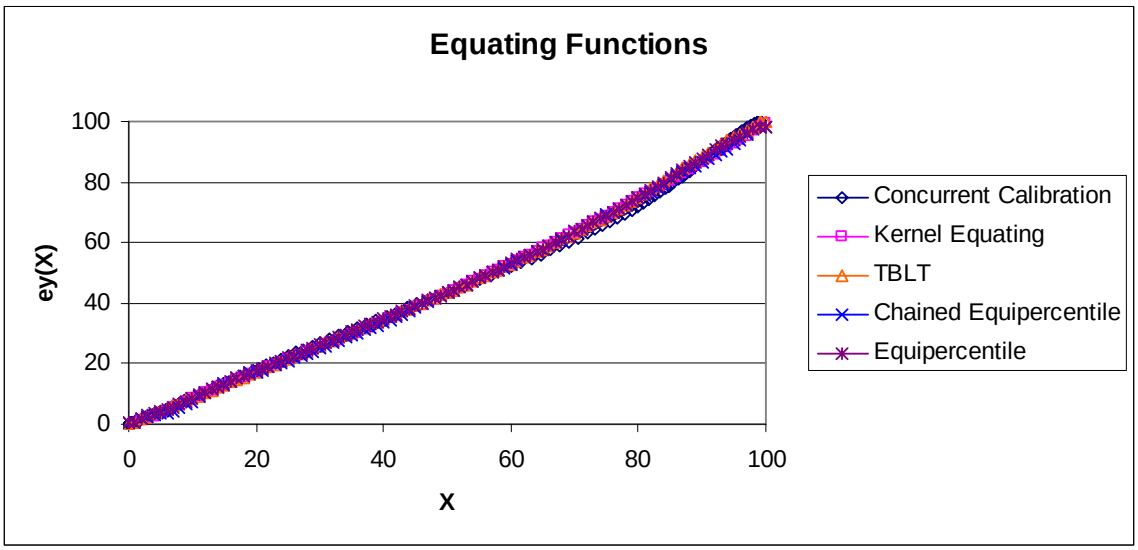


Figure E.21: 100 Items per form, 35% Anchor Length, 100,000 Sample Size, No Ability Difference

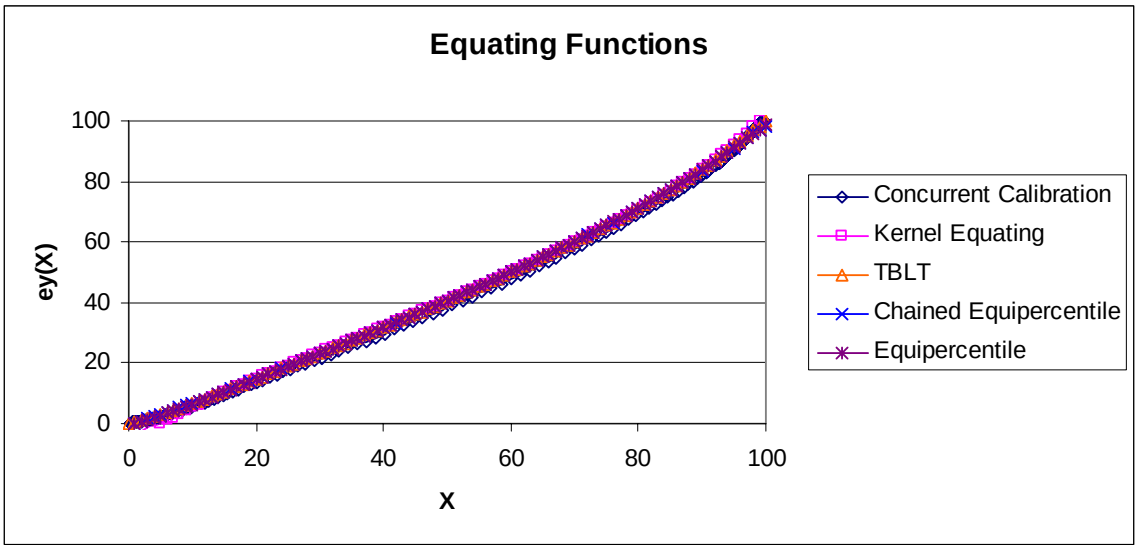


Figure E.22: 100 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

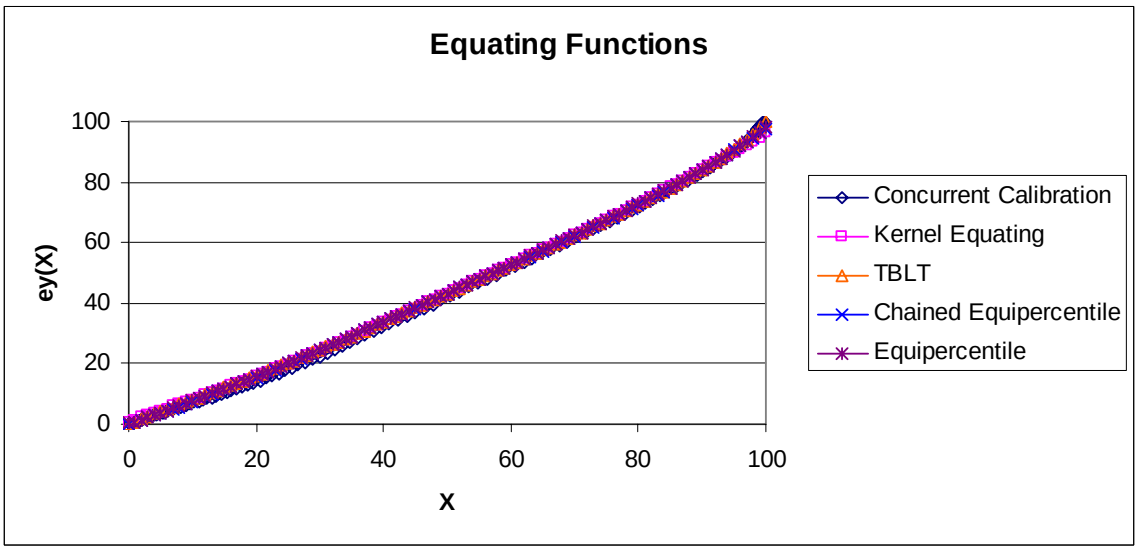


Figure E.23: 100 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

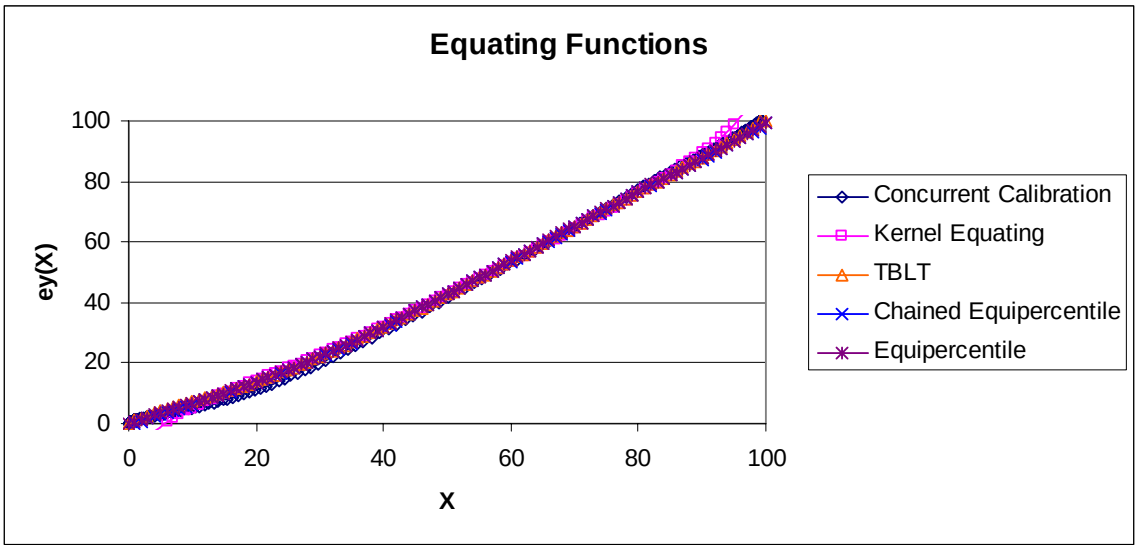


Figure E.24: 100 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

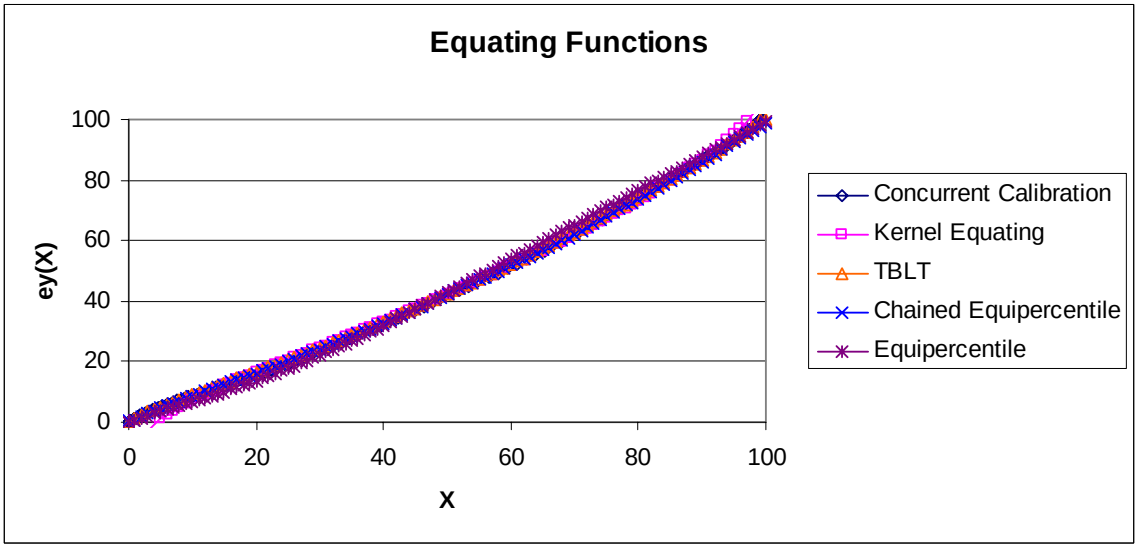


Figure E.25: 100 Items per form, 20% Anchor Length, 1000 Sample Size, No Ability Difference

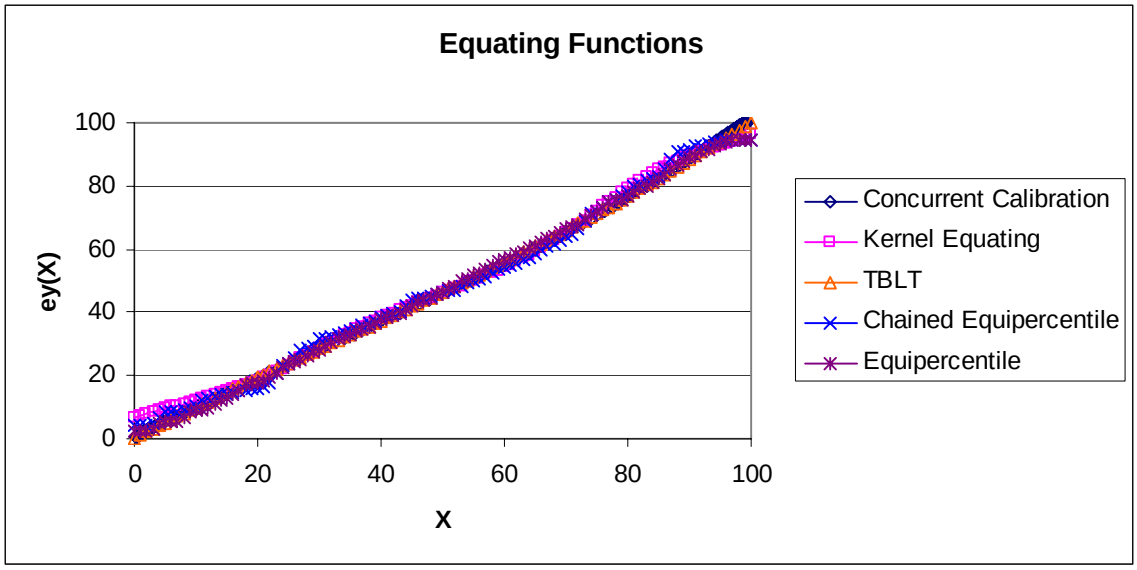


Figure E.26: 100 Items per form, 20% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

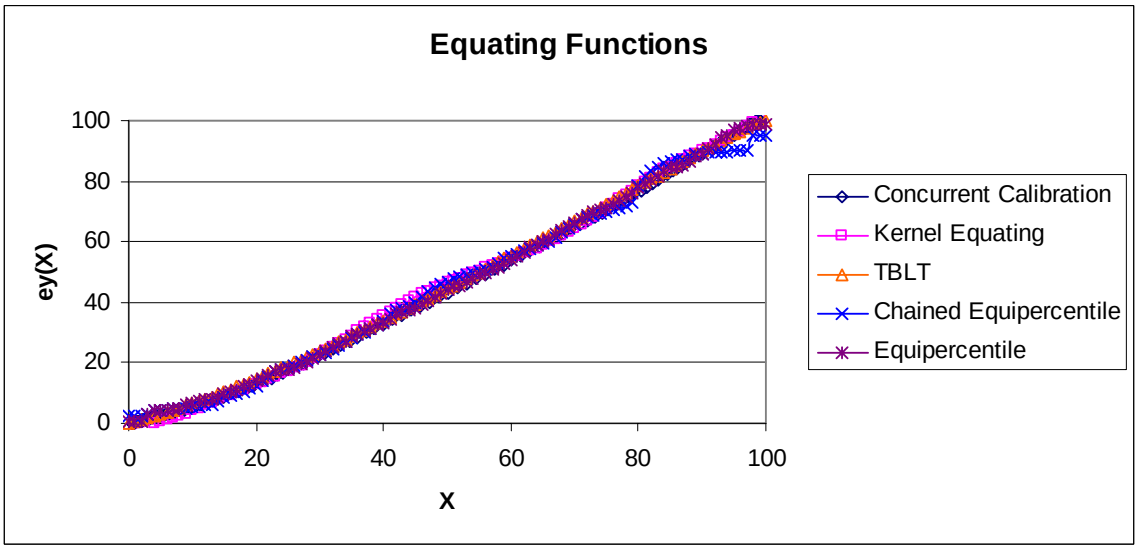


Figure E.27: 100 Items per form, 20% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

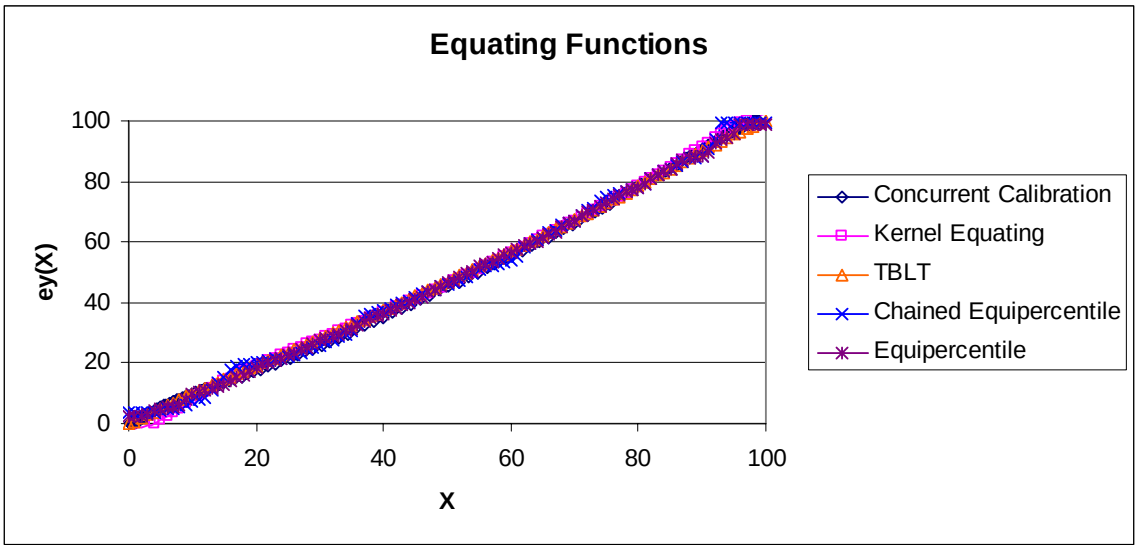


Figure E.28: 100 Items per form, 20% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

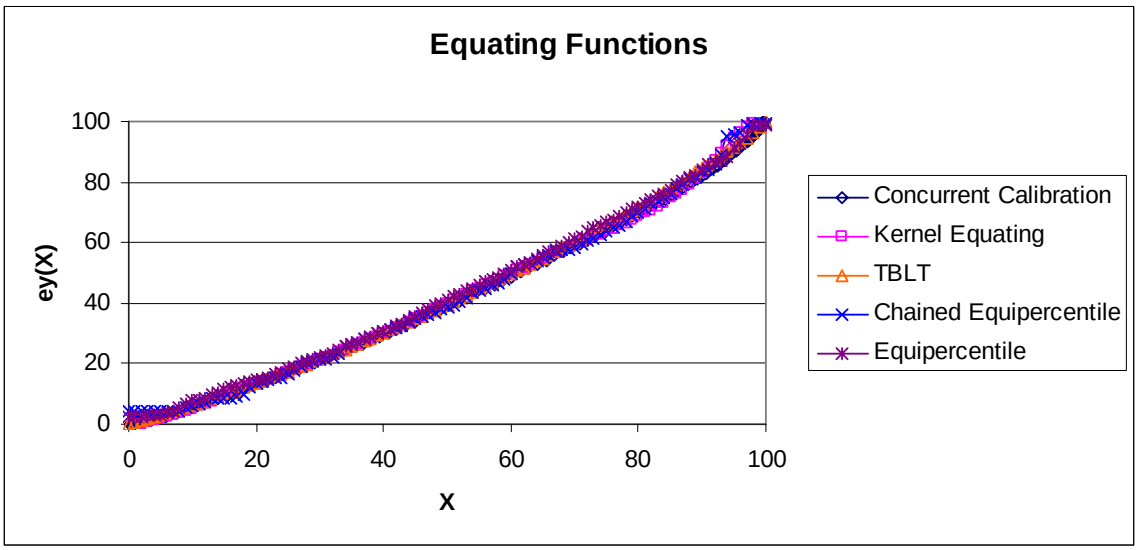


Figure E.29: 100 Items per form, 20% Anchor Length, 10,000 Sample Size, No Ability Difference

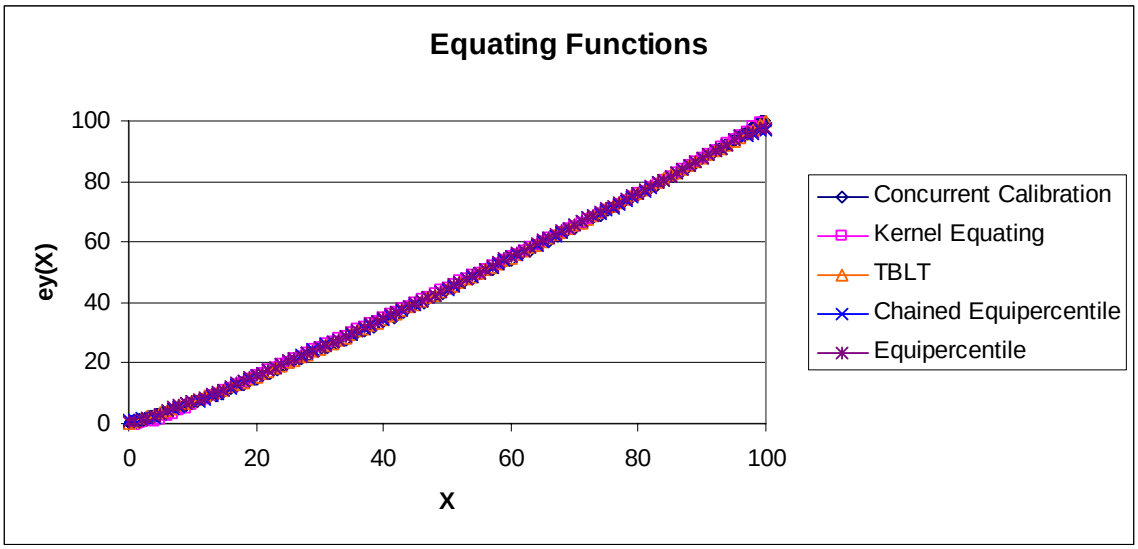


Figure E.30: 100 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

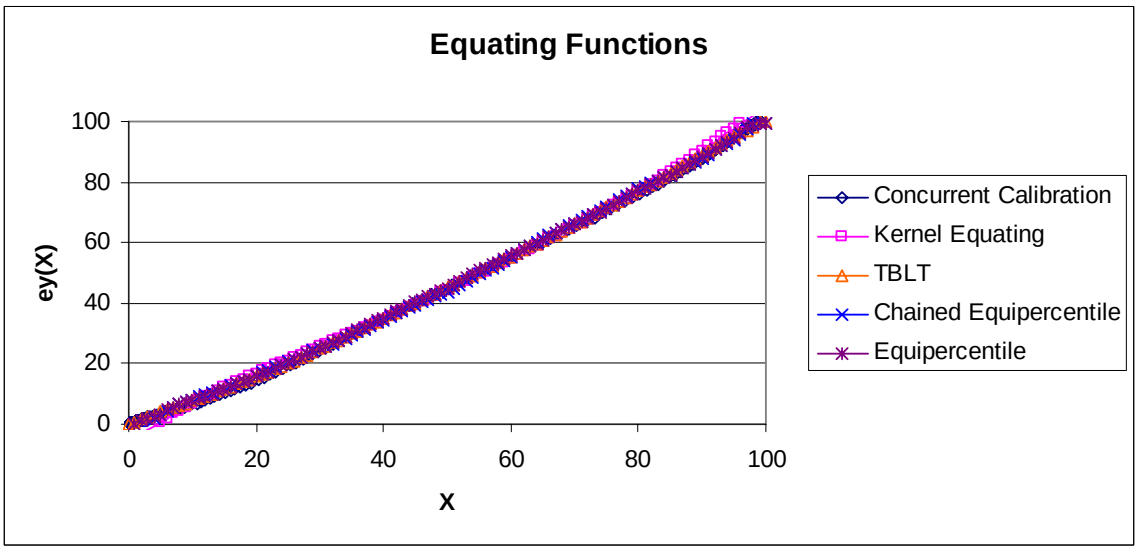


Figure E.31: 100 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

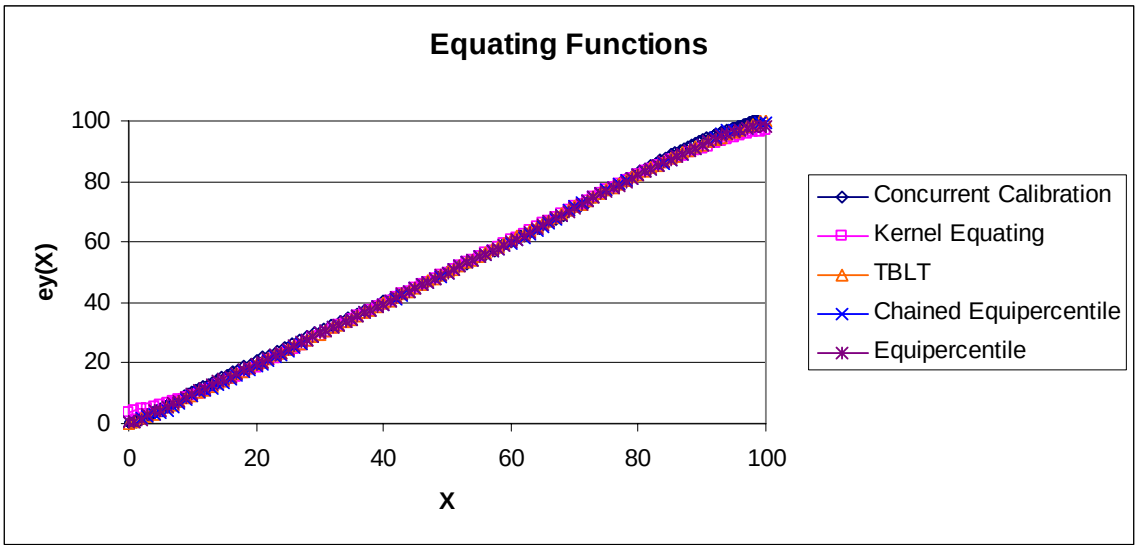


Figure E.32: 100 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

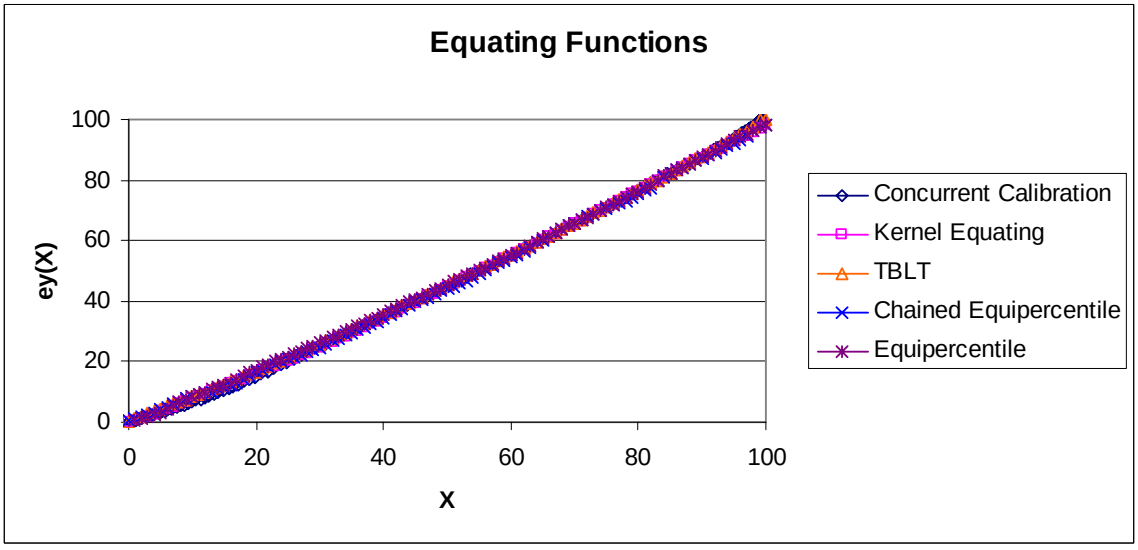


Figure E.33: 100 Items per form, 20% Anchor Length, 100,000 Sample Size, No Ability Difference

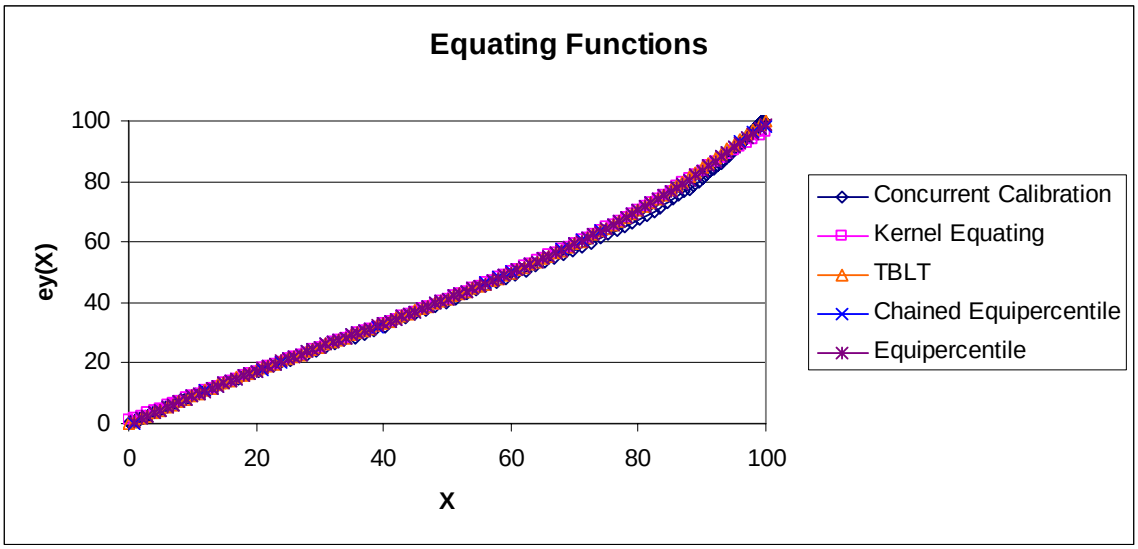


Figure E.34: 100 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

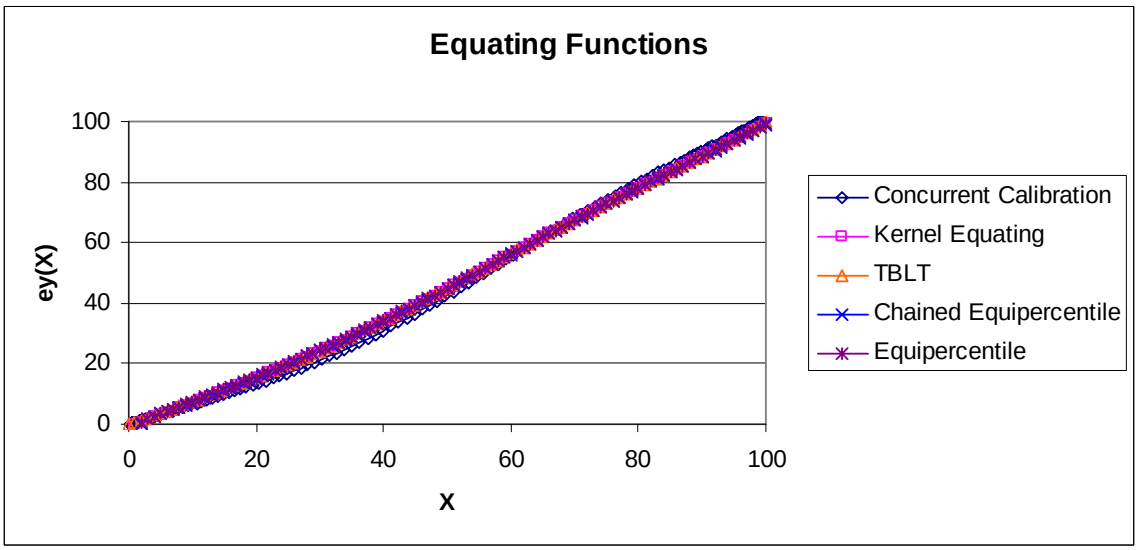


Figure E.35: 100 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

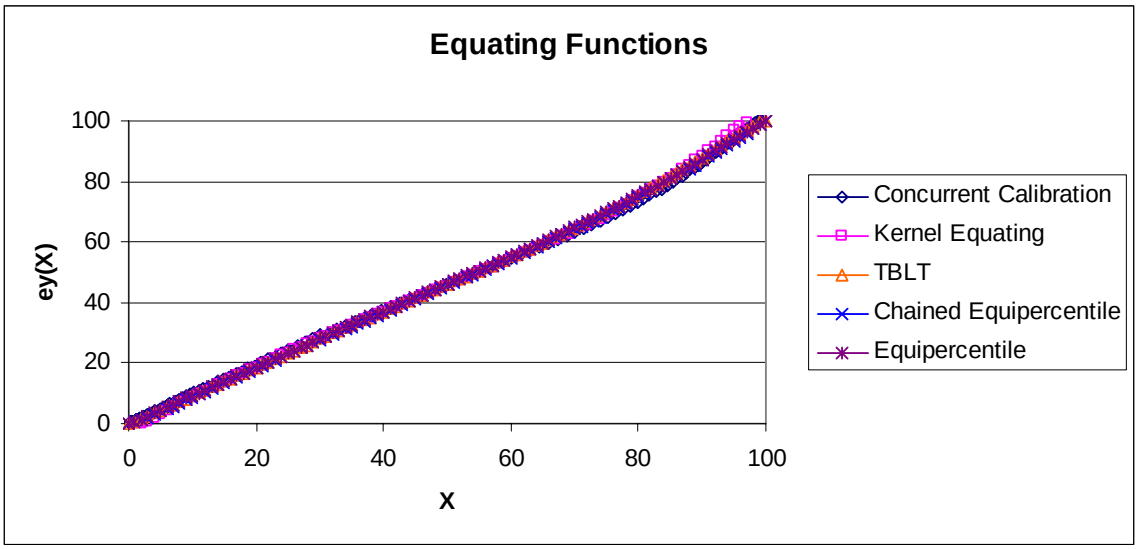


Figure E.36: 100 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

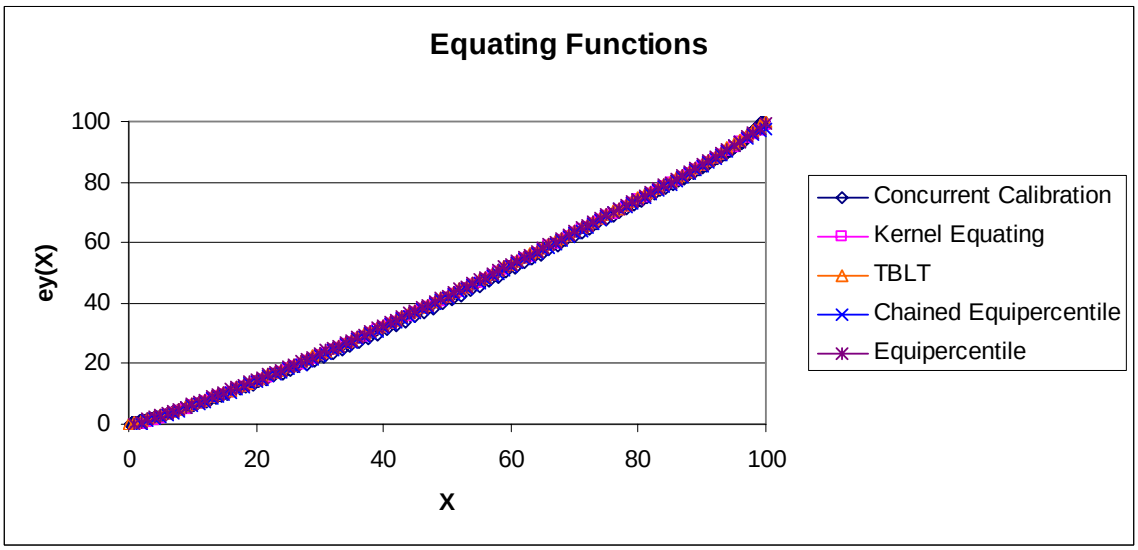


Figure E.37: 60 Items per form, 50% Anchor Length, 1000 Sample Size, No Ability Difference

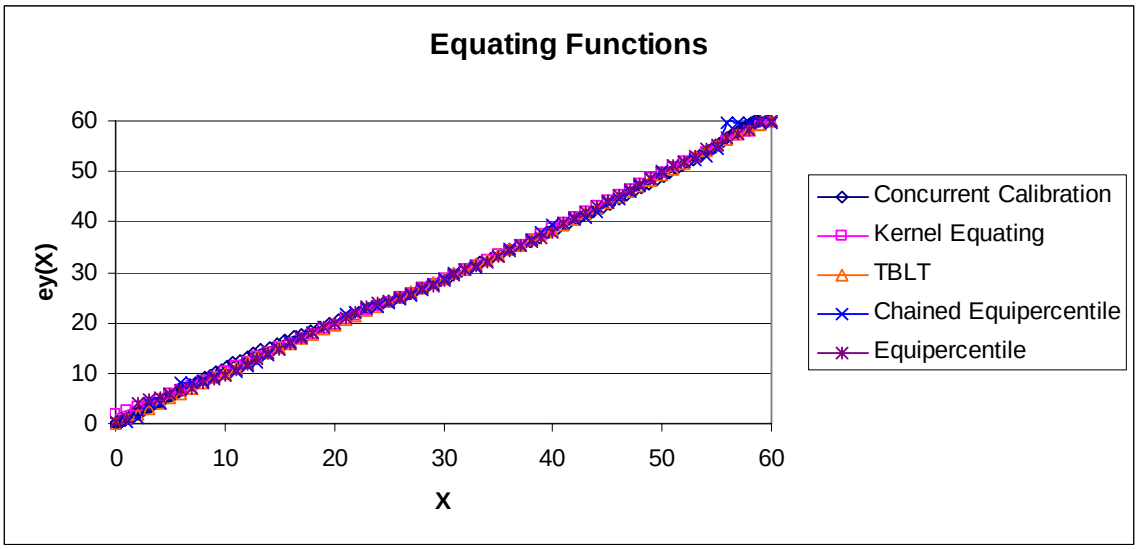


Figure E.38: 60 Items per form, 50% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

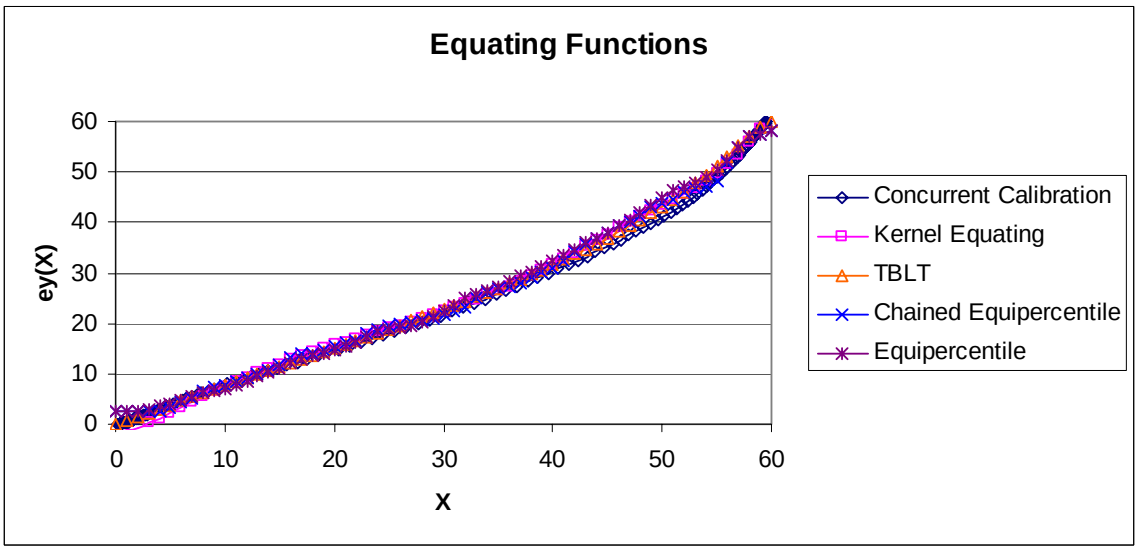


Figure E.39: 60 Items per form, 50% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

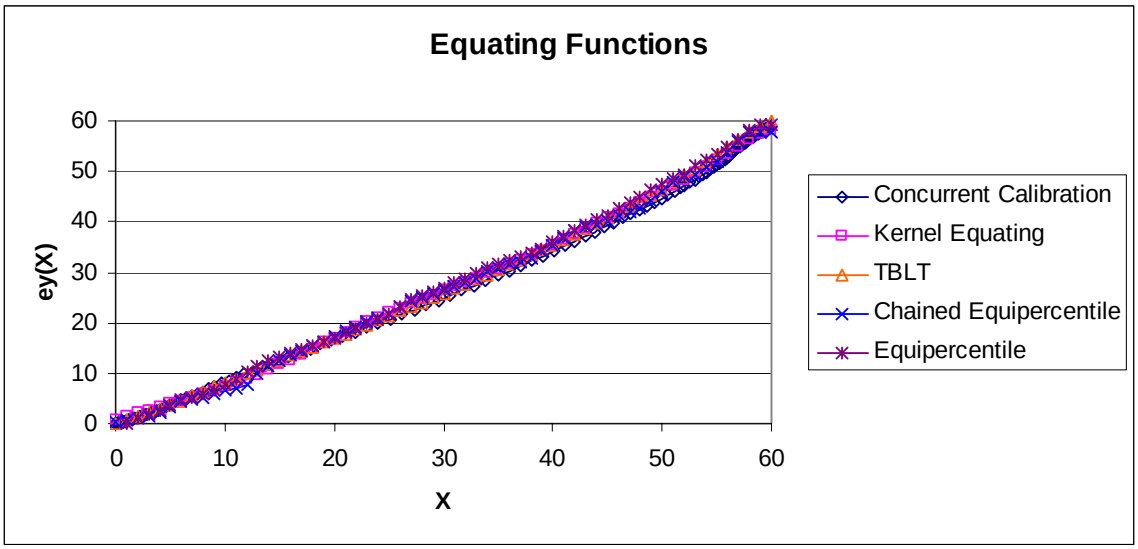


Figure E.40: 60 Items per form, 50% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

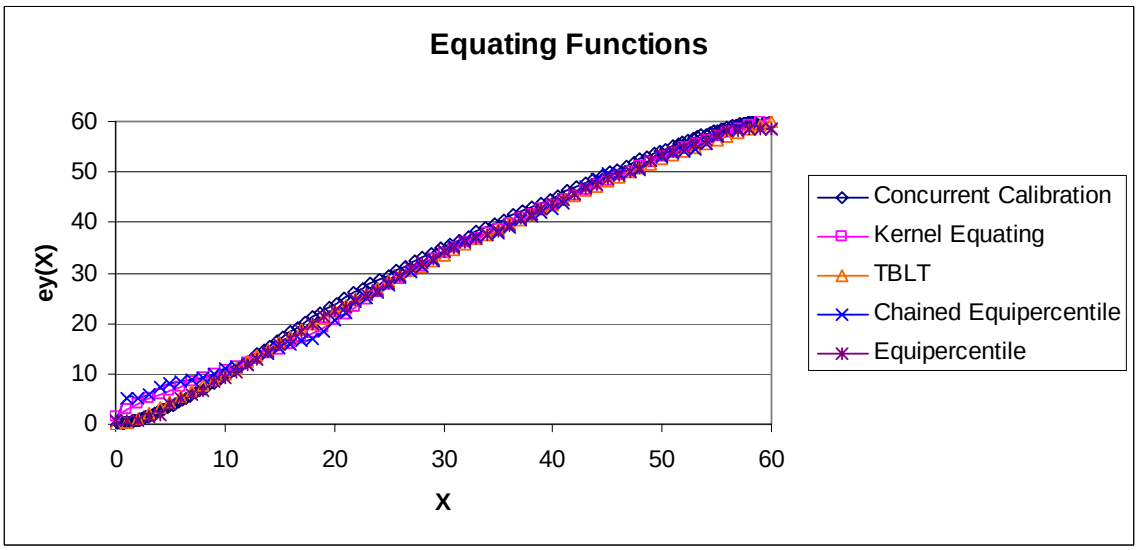


Figure E.41: 60 Items per form, 50% Anchor Length, 10,000 Sample Size, No Ability Difference

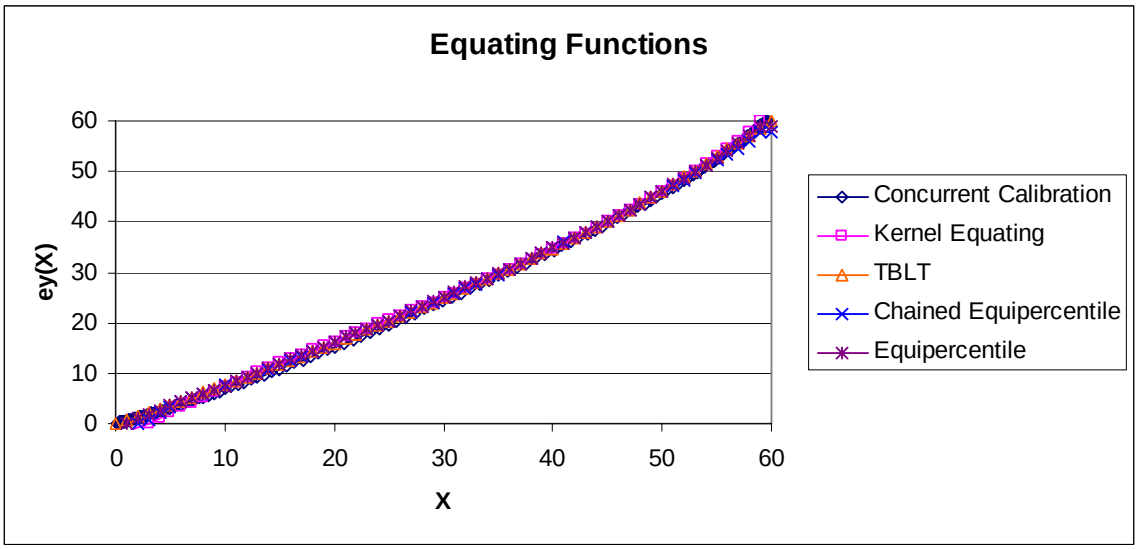


Figure E.42: 60 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

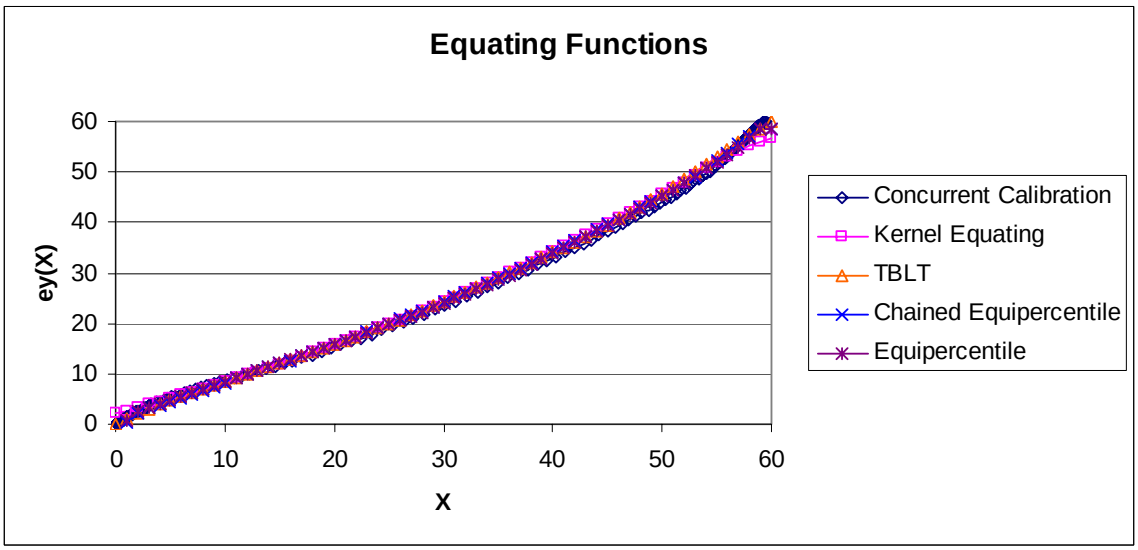


Figure E.43: 60 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

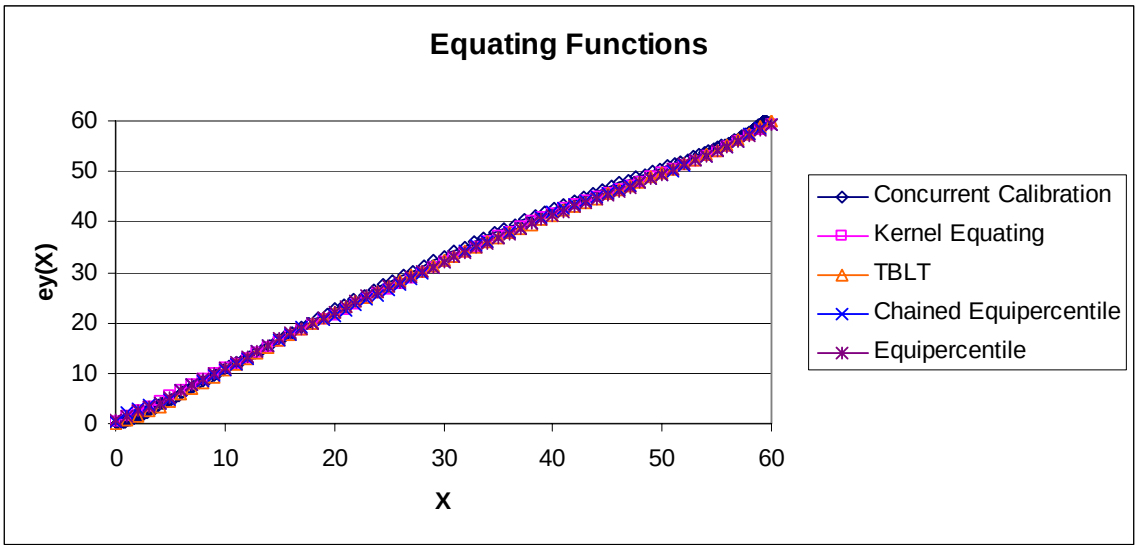


Figure E.44: 60 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

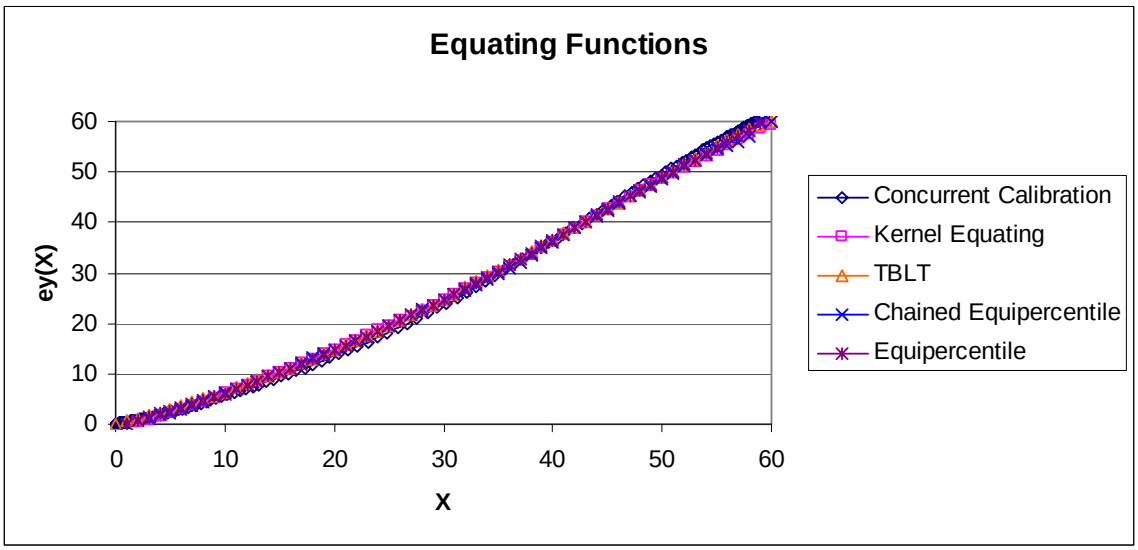


Figure E.45: 60 Items per form, 50% Anchor Length, 100,000 Sample Size, No Ability Difference

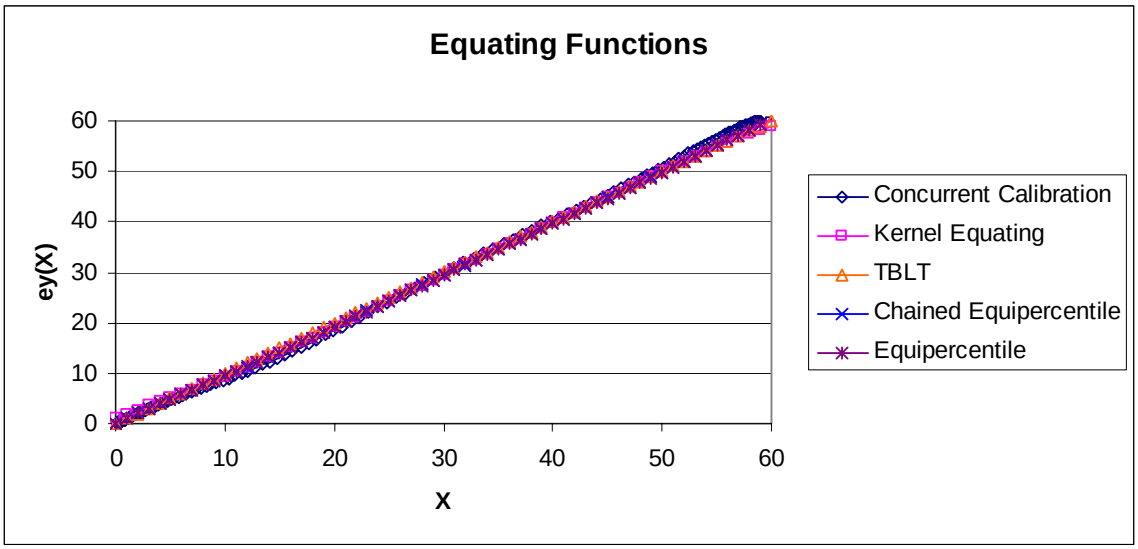


Figure E.46: 60 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

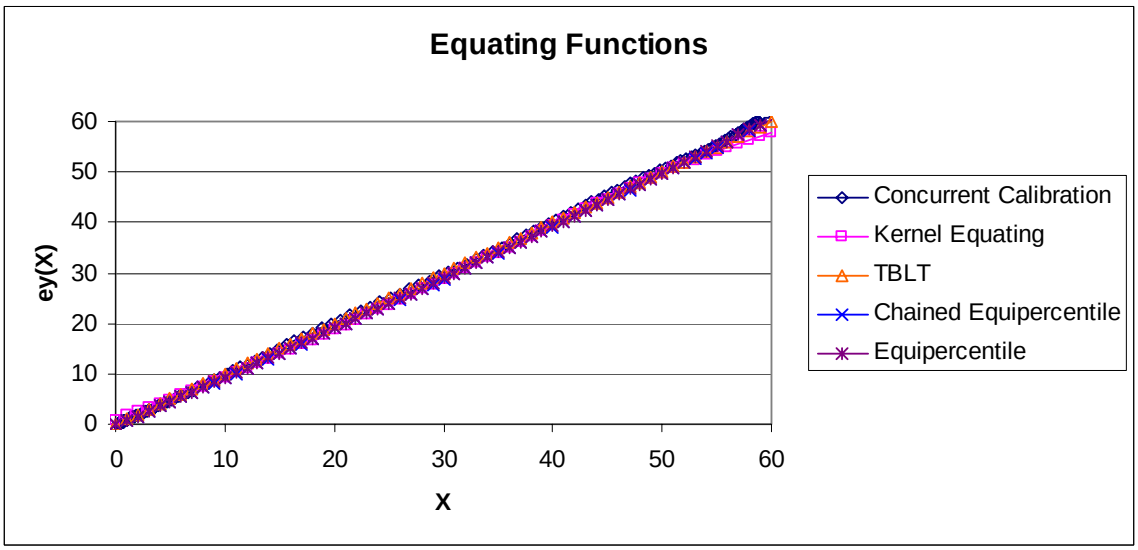


Figure E.47: 60 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

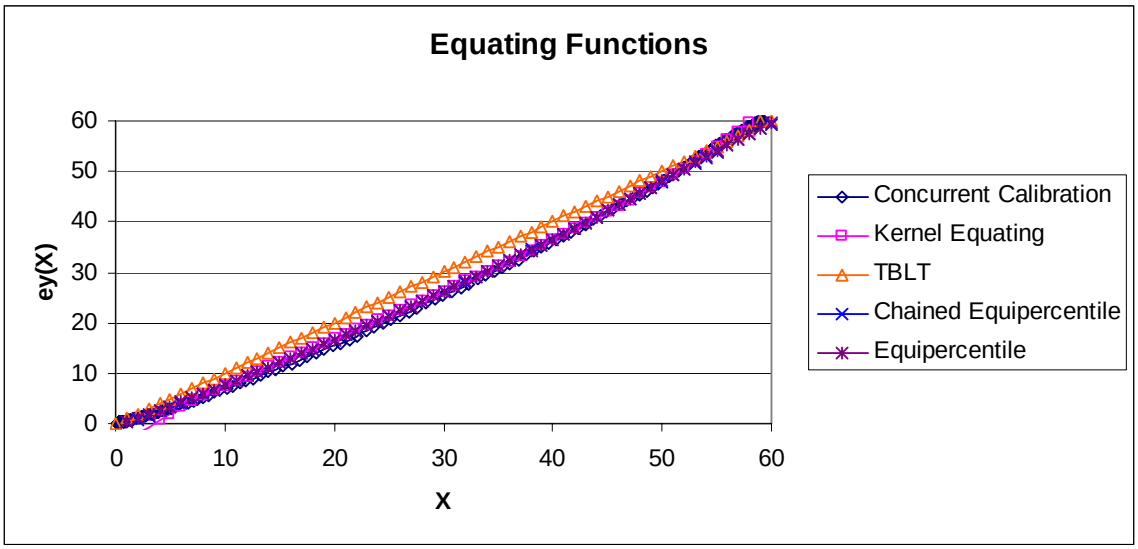


Figure E.48: 60 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

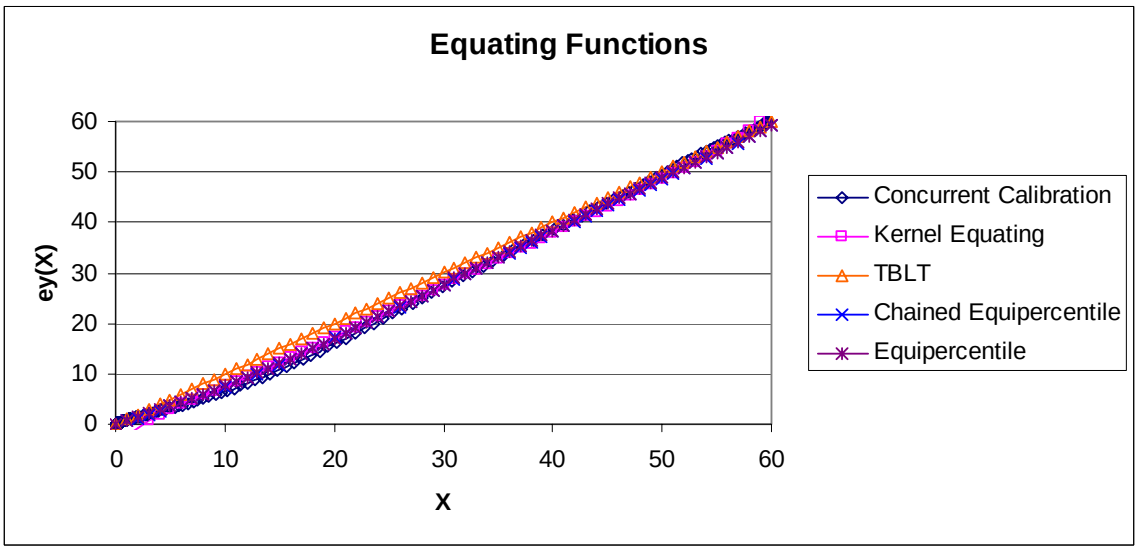


Figure E.49: 60 Items per form, 35% Anchor Length, 1000 Sample Size, No Ability Difference

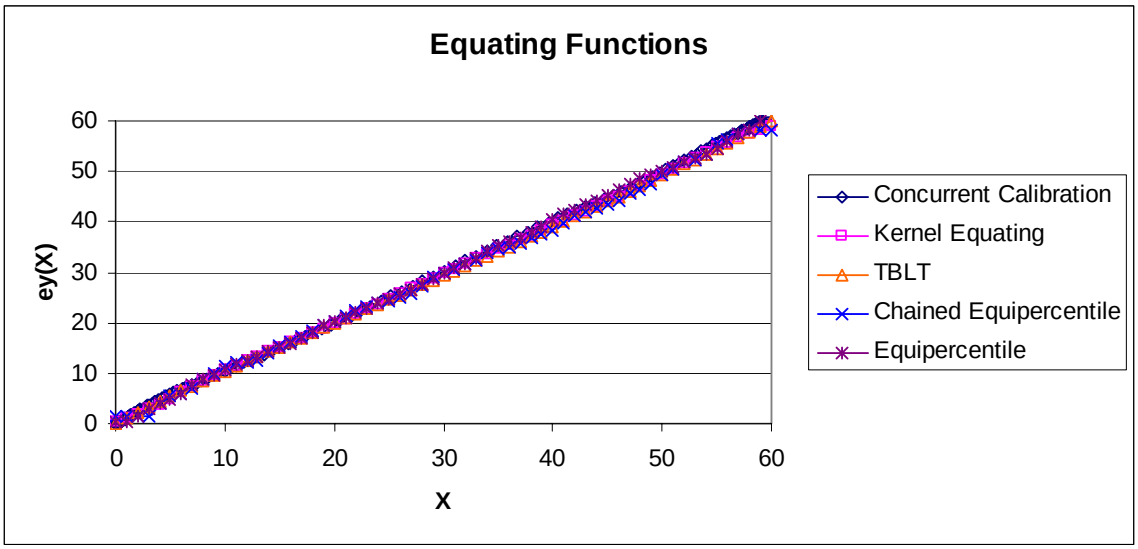


Figure E.50: 60 Items per form, 35% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

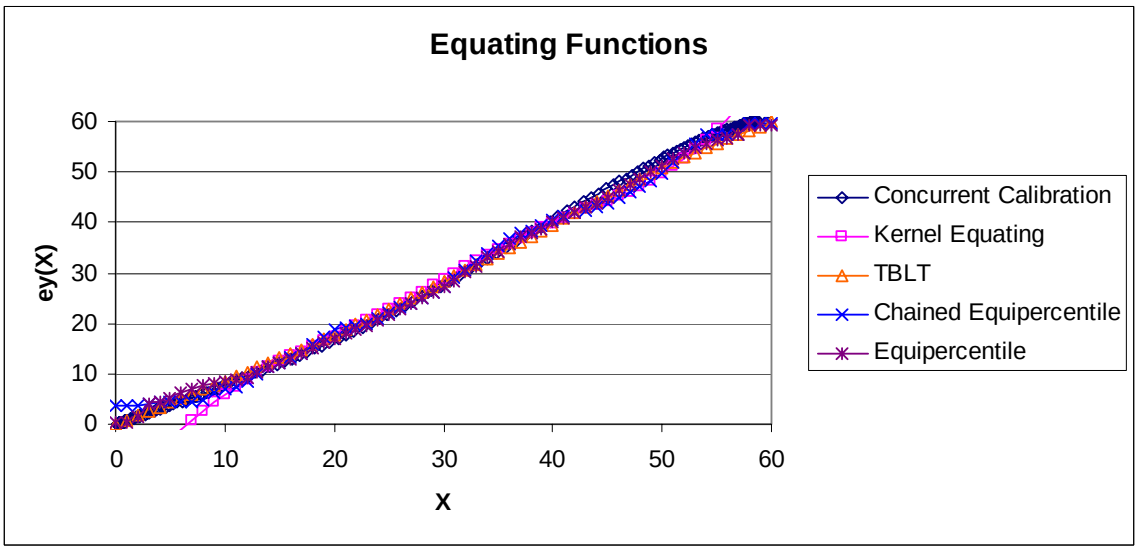


Figure E.51: 60 Items per form, 35% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

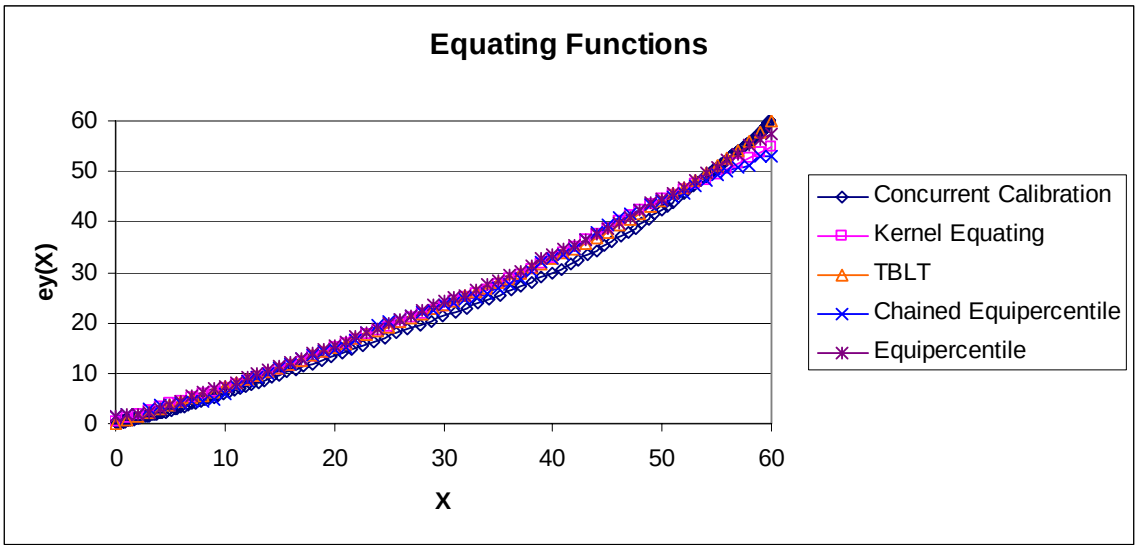


Figure E.52: 60 Items per form, 35% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

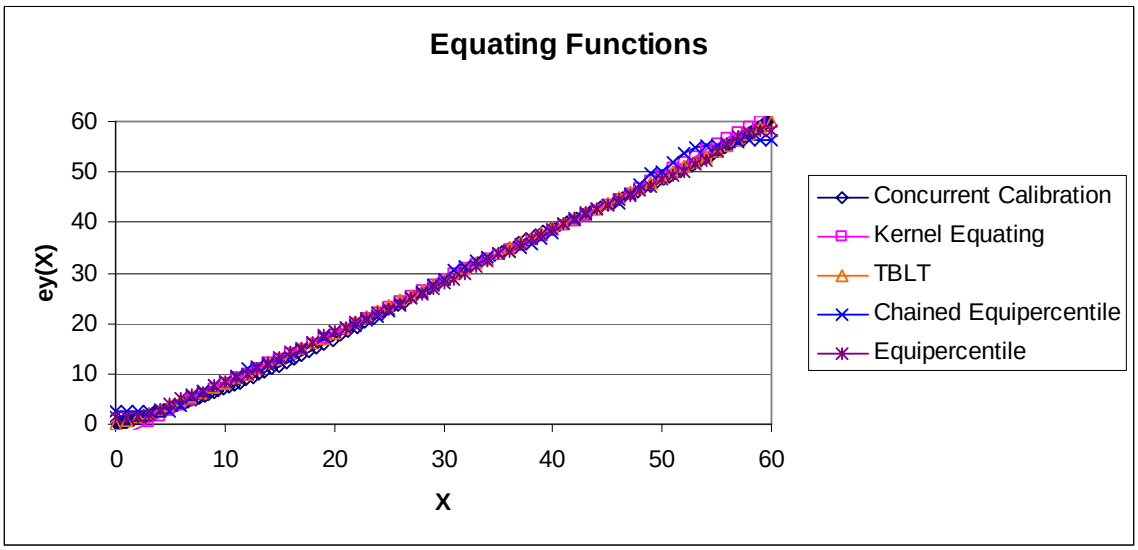


Figure E.53: 60 Items per form, 35% Anchor Length, 10,000 Sample Size, No Ability Difference

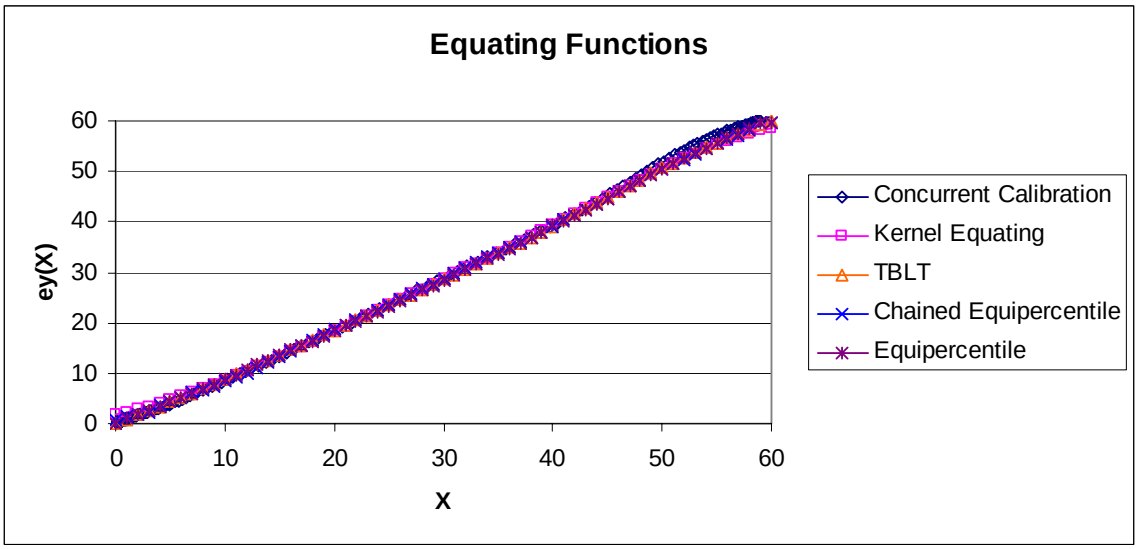


Figure E.54: 60 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

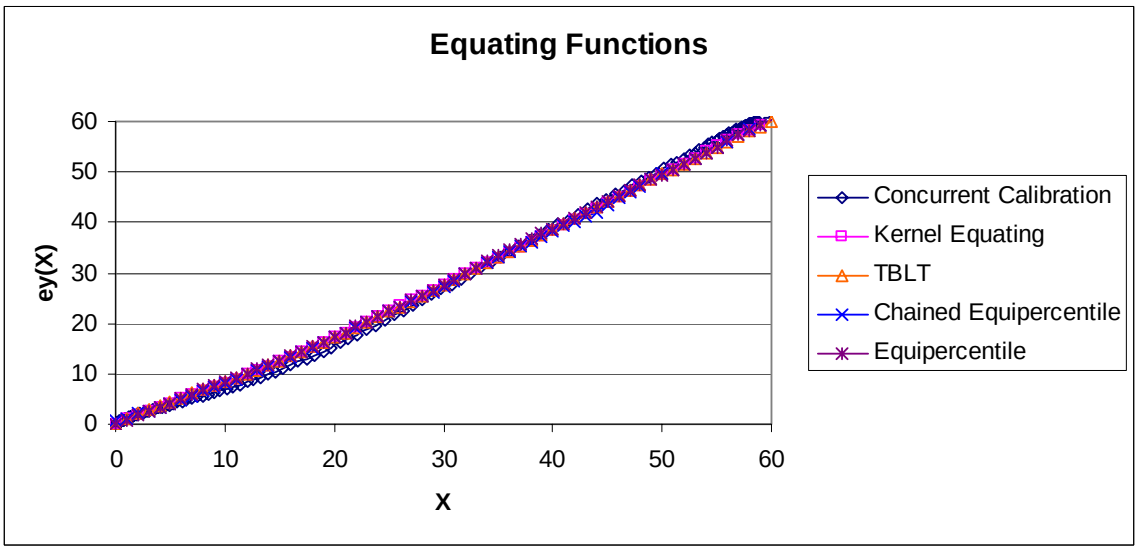


Figure E.55: 60 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

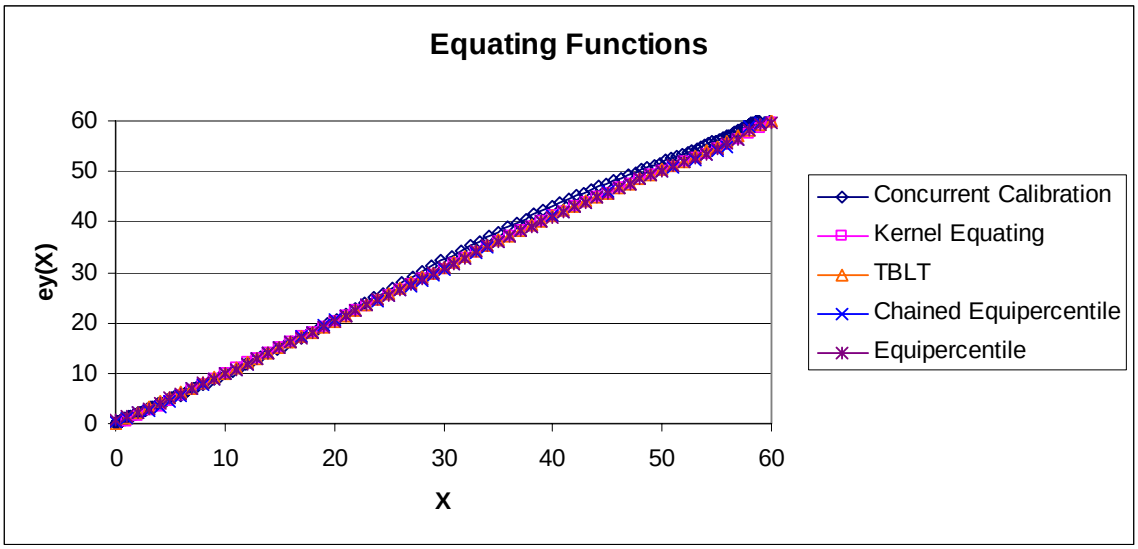


Figure E.56: 60 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

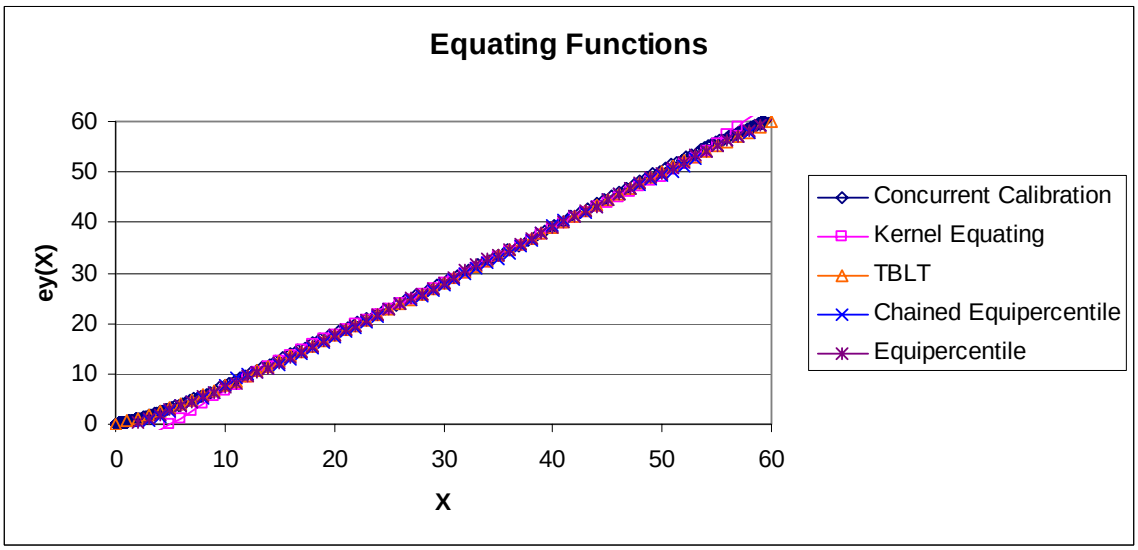


Figure E.57: 60 Items per form, 35% Anchor Length, 100,000 Sample Size, No Ability Difference

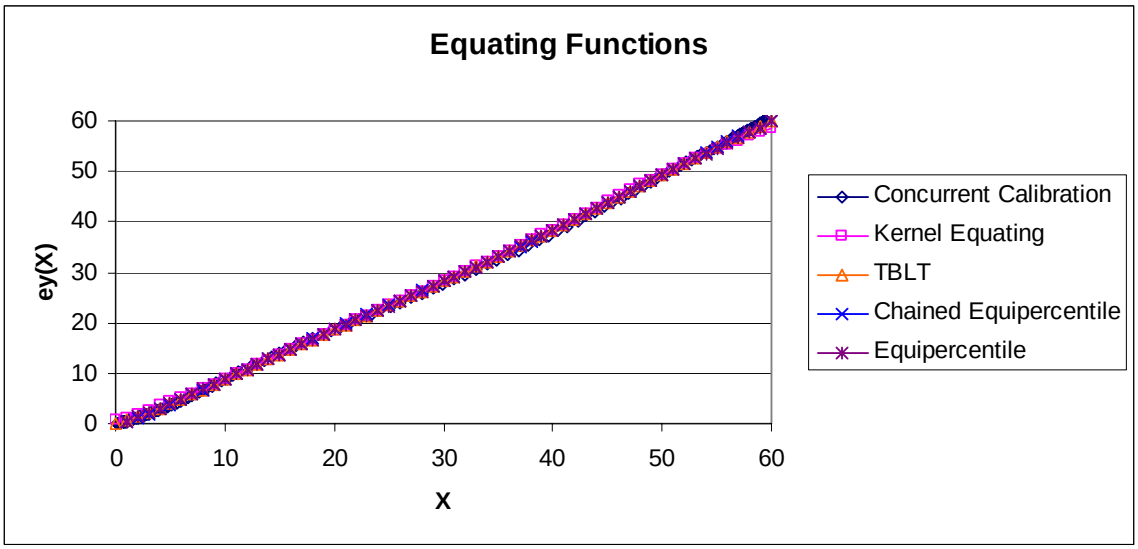


Figure E.58: 60 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

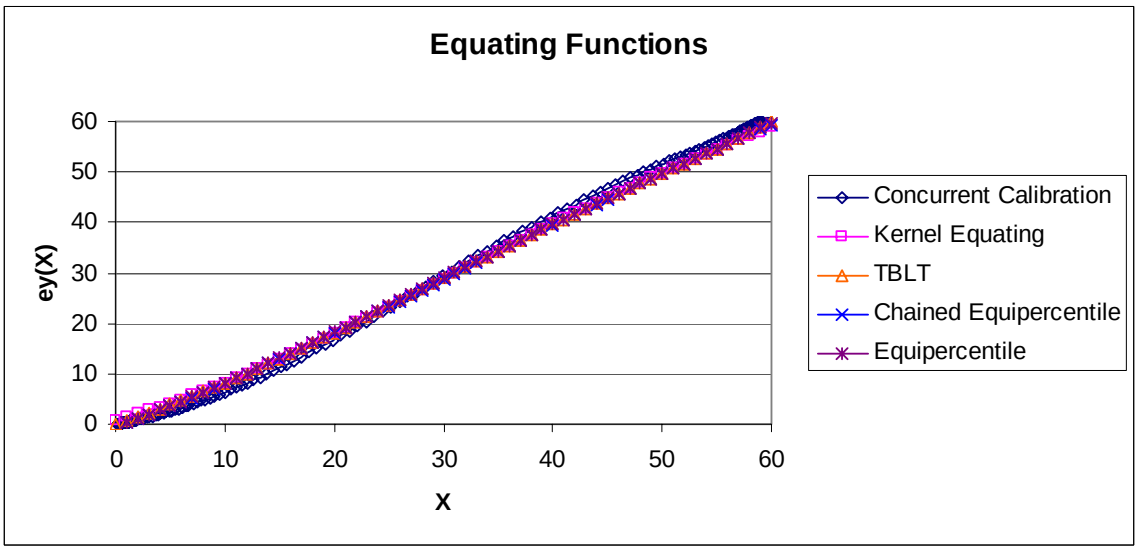


Figure E.59: 60 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

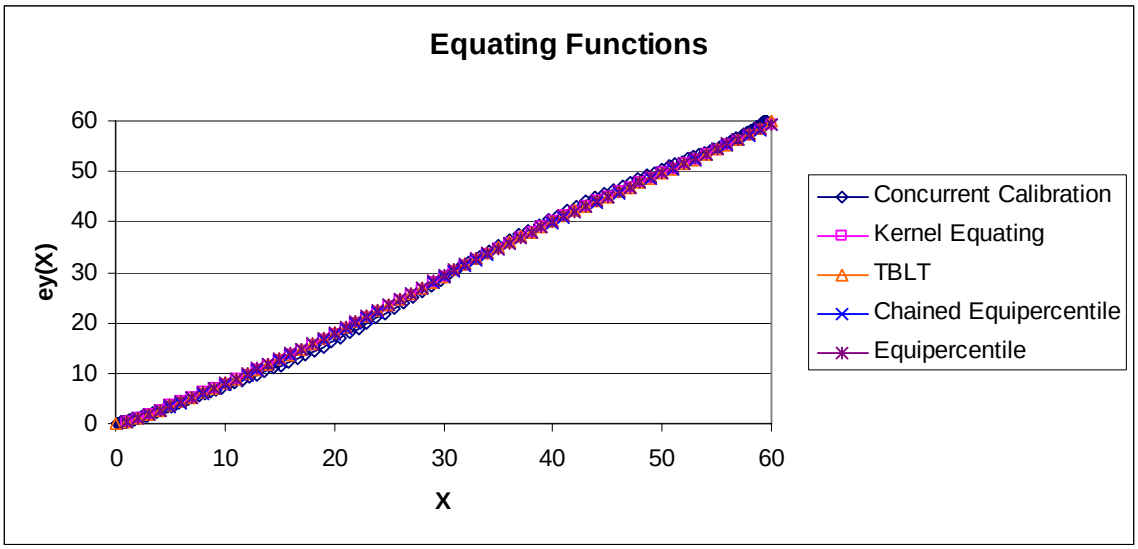


Figure E.60: 60 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

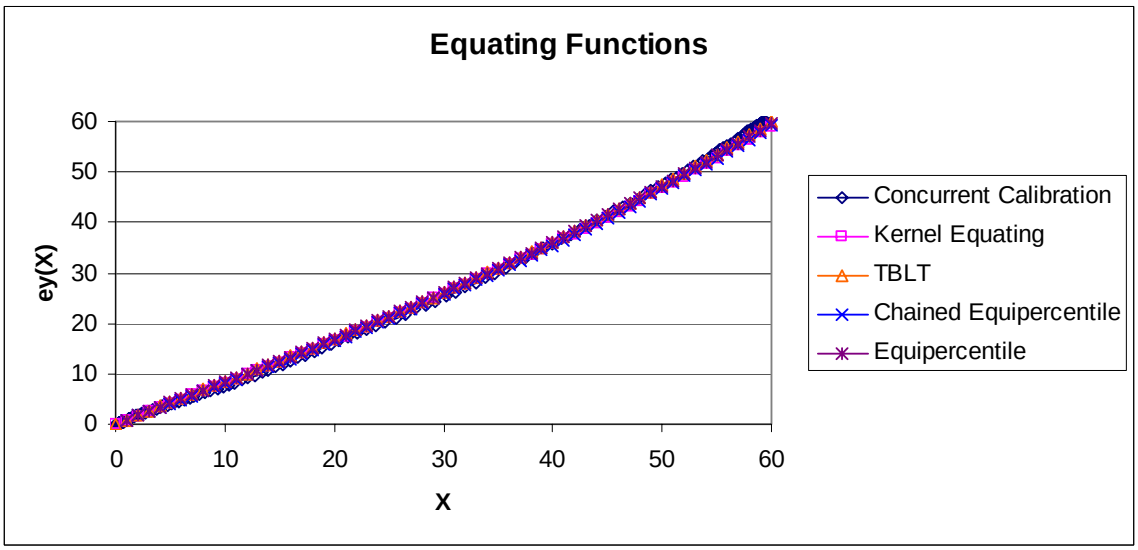


Figure E.61: 60 Items per form, 20% Anchor Length, 1000 Sample Size, No Ability Difference

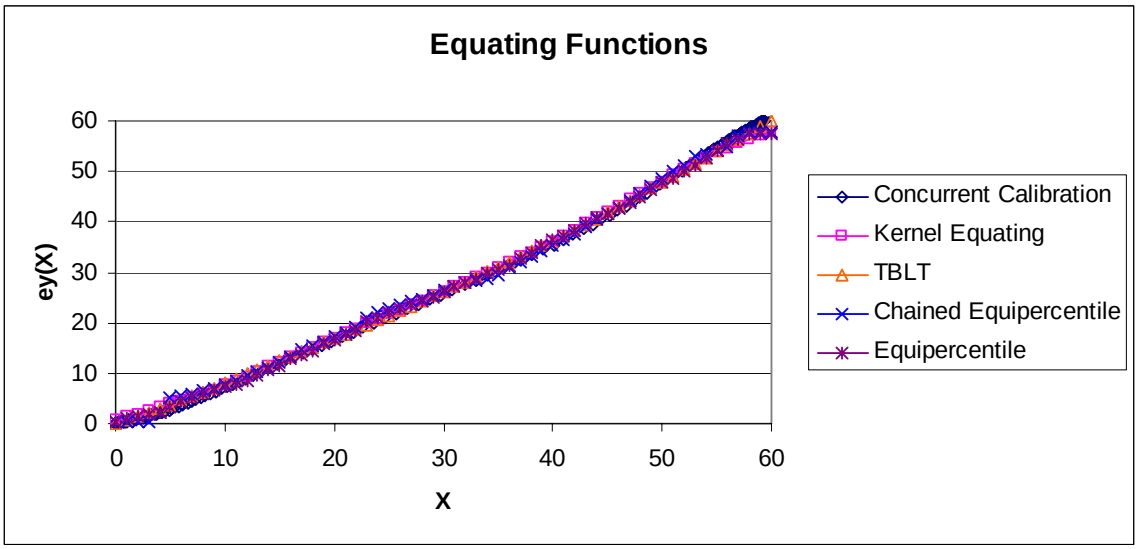


Figure E.62: 60 Items per form, 20% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

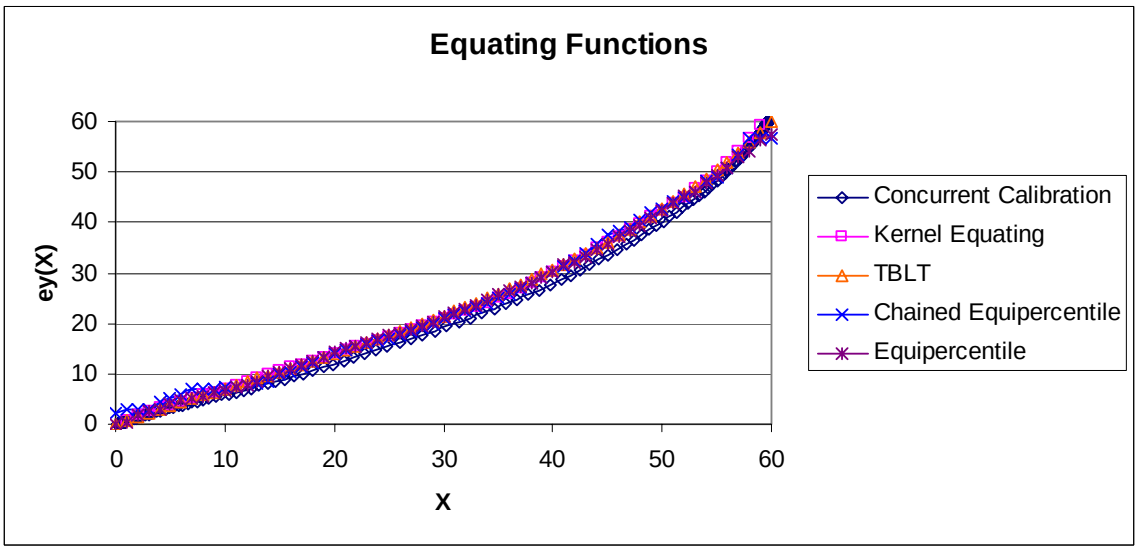


Figure E.63: 60 Items per form, 20% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

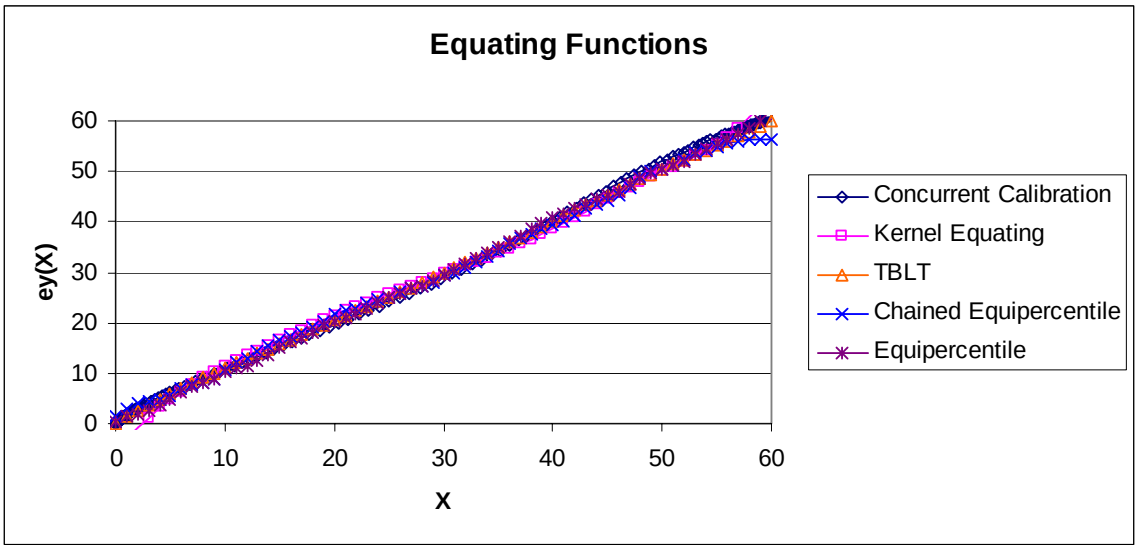


Figure E.64: 60 Items per form, 20% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

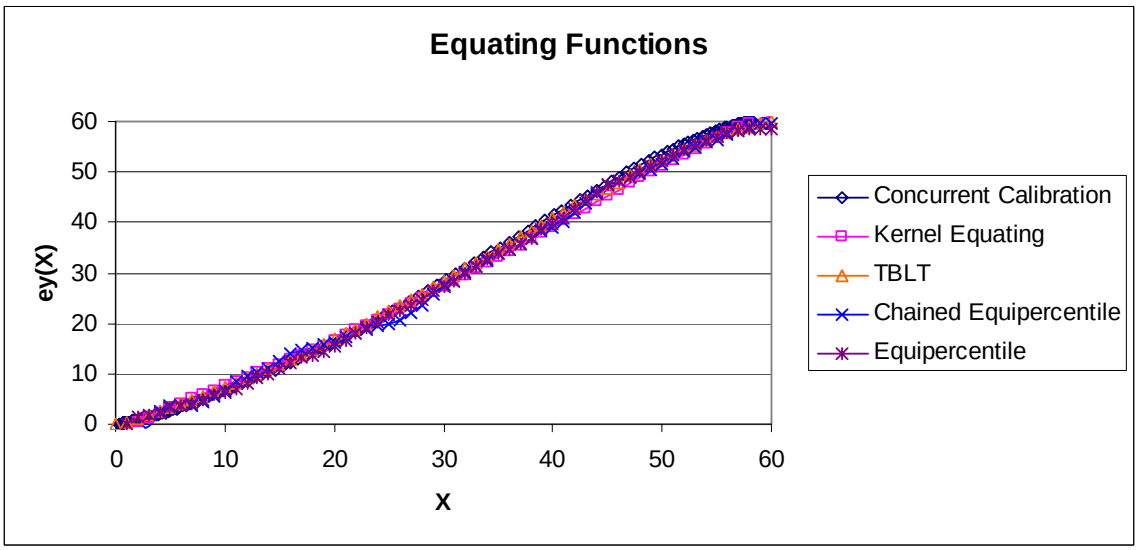


Figure E.65: 60 Items per form, 20% Anchor Length, 10,000 Sample Size, No Ability Difference

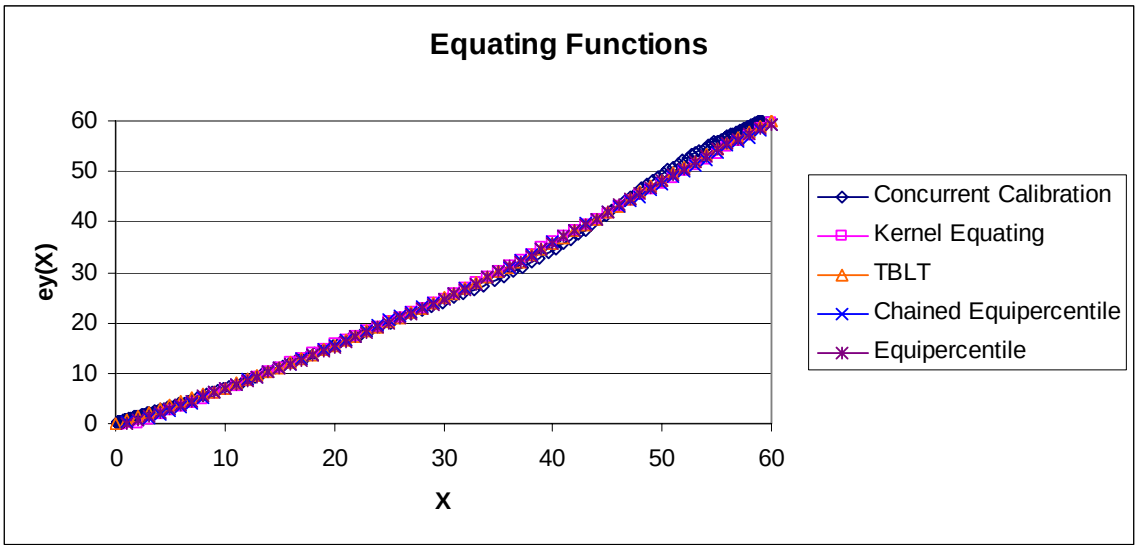


Figure E.66: 60 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

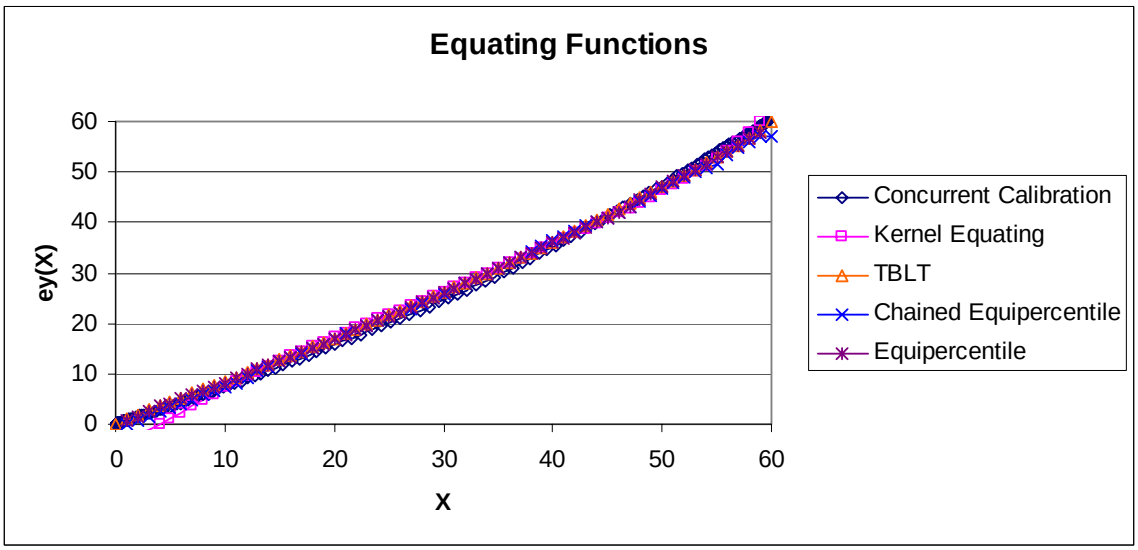


Figure E.67: 60 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

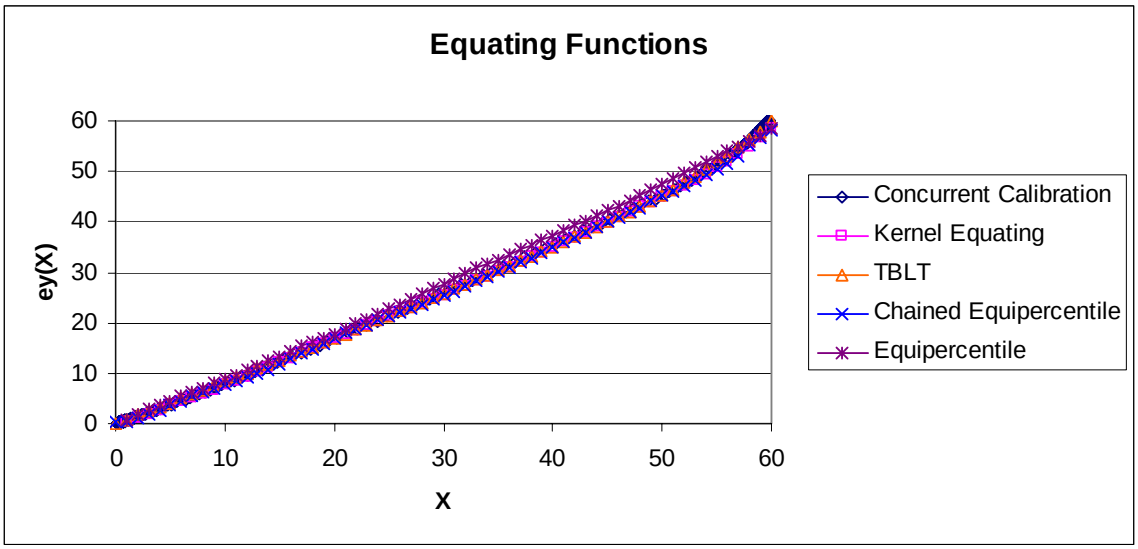


Figure E.68: 60 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

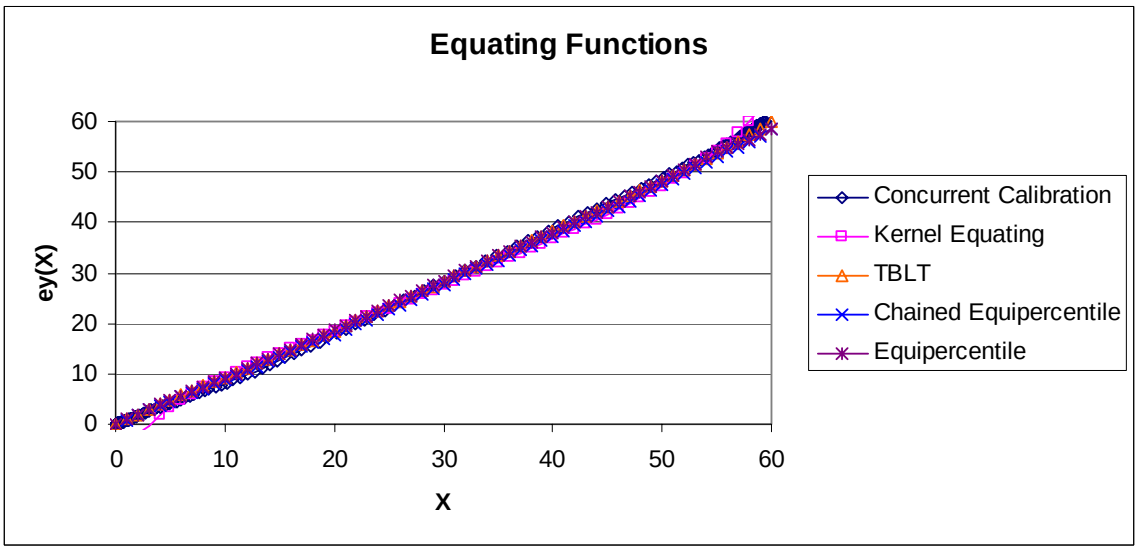


Figure E.69: 60 Items per form, 20% Anchor Length, 100,000 Sample Size, No Ability Difference

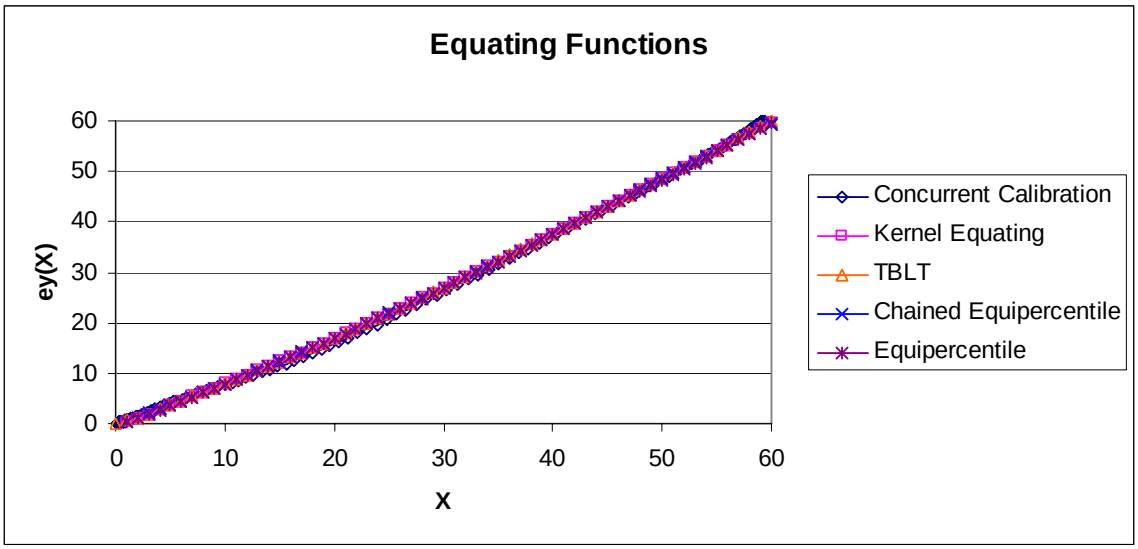


Figure E.70: 60 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

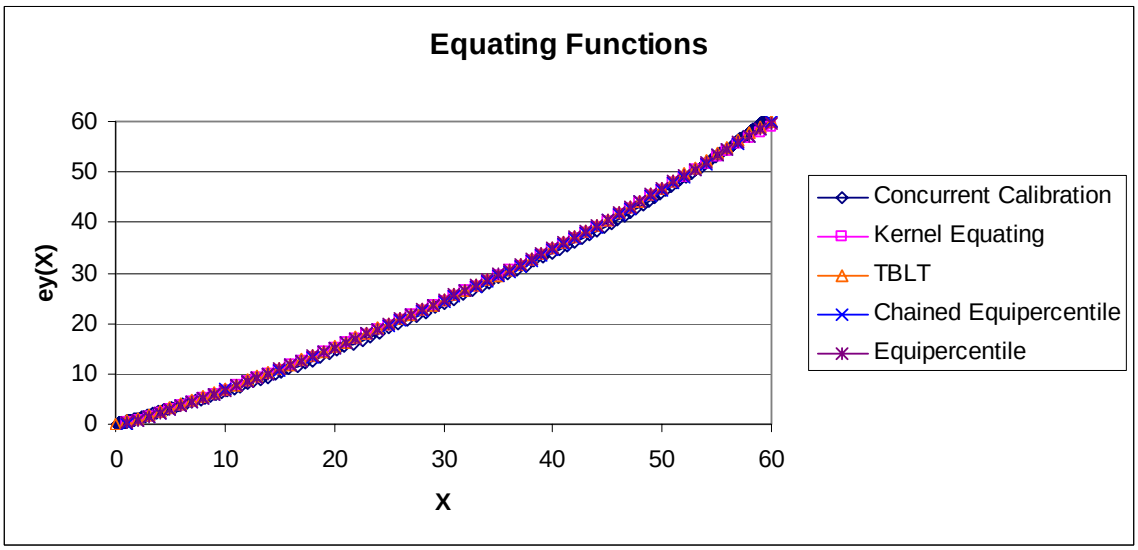


Figure E.71: 60 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

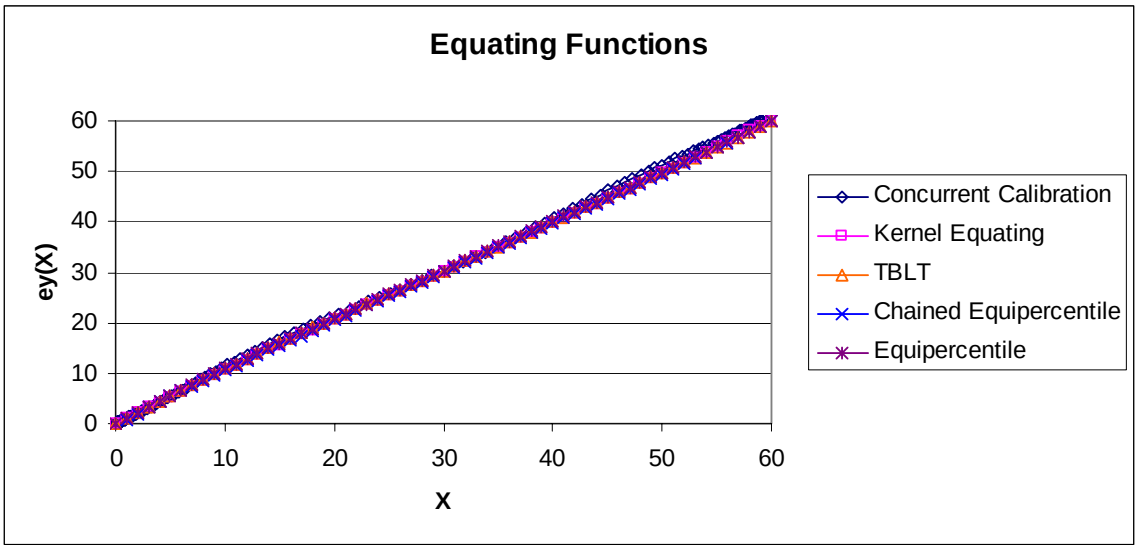


Figure E.72: 60 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

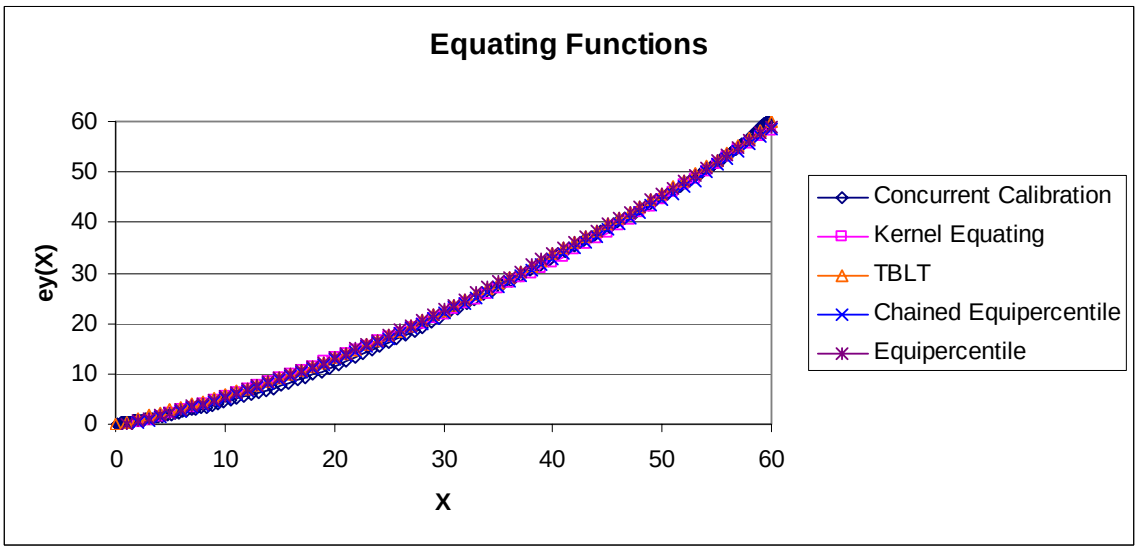


Figure E.73: 20 Items per form, 50% Anchor Length, 1000 Sample Size, No Ability Difference

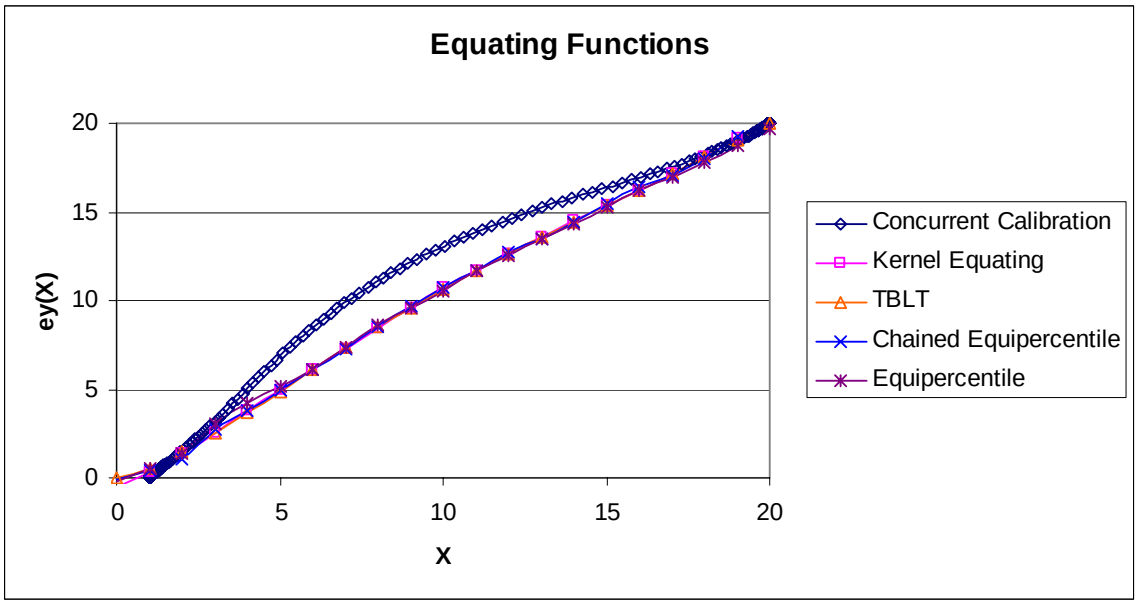


Figure E.74: 20 Items per form, 50% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

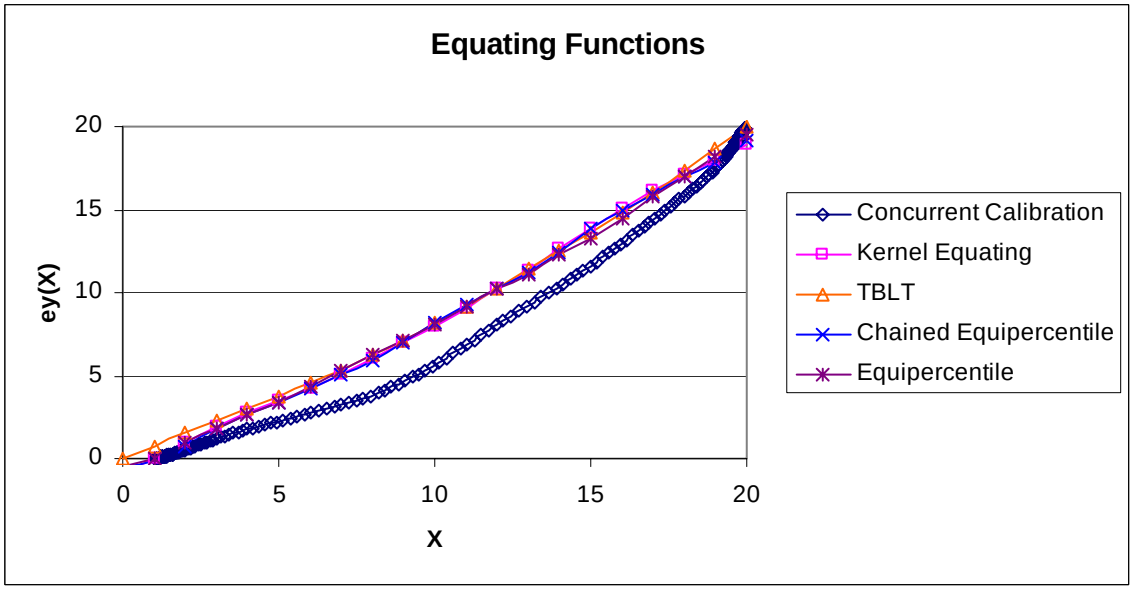


Figure E.75: 20 Items per form, 50% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

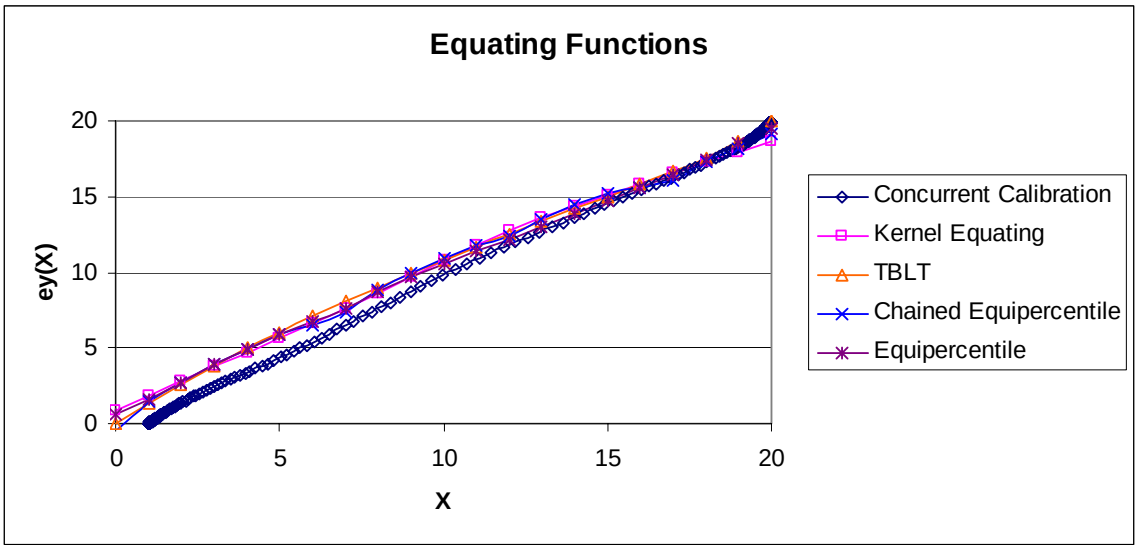


Figure E.76: 20 Items per form, 50% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

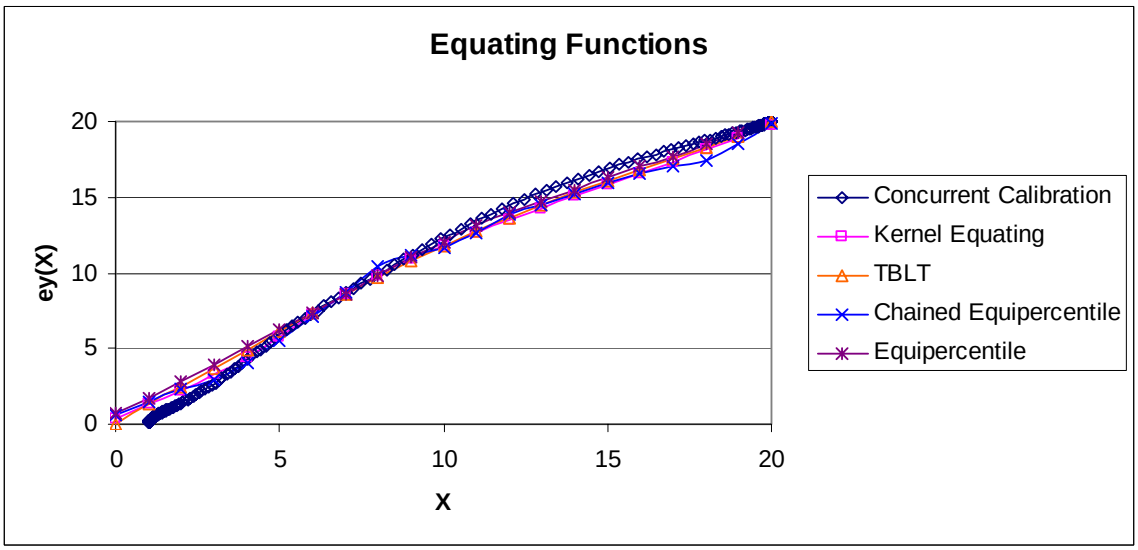


Figure E.77: 20 Items per form, 50% Anchor Length, 10,000 Sample Size, No Ability Difference

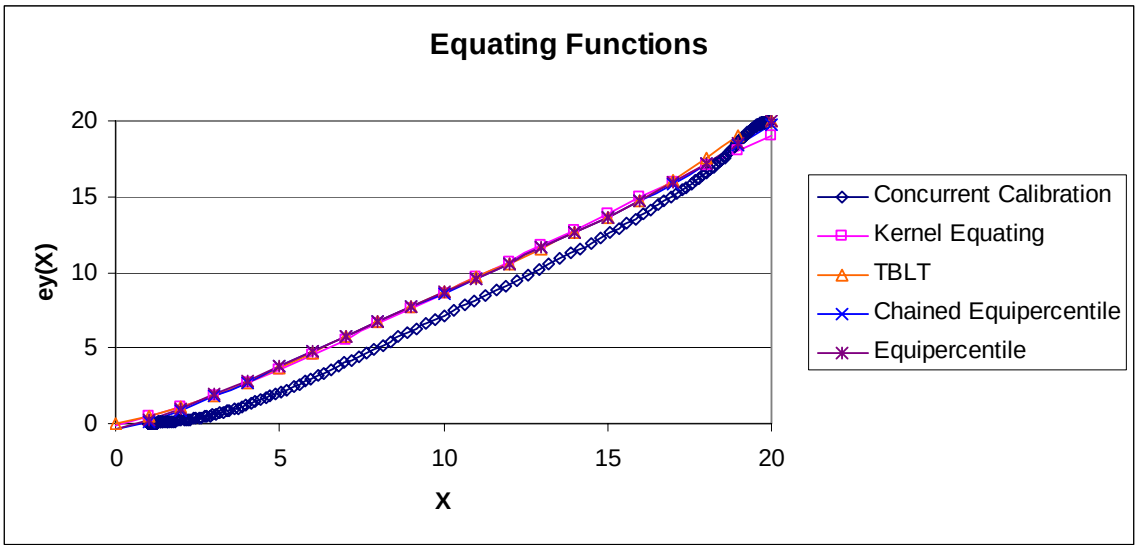


Figure E.78: 20 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

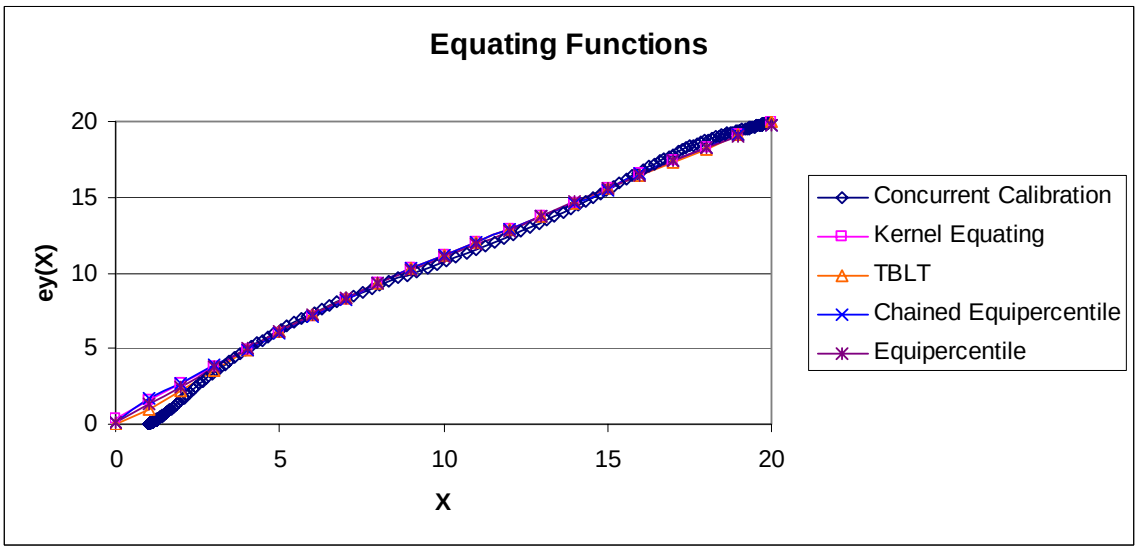


Figure E.79: 20 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

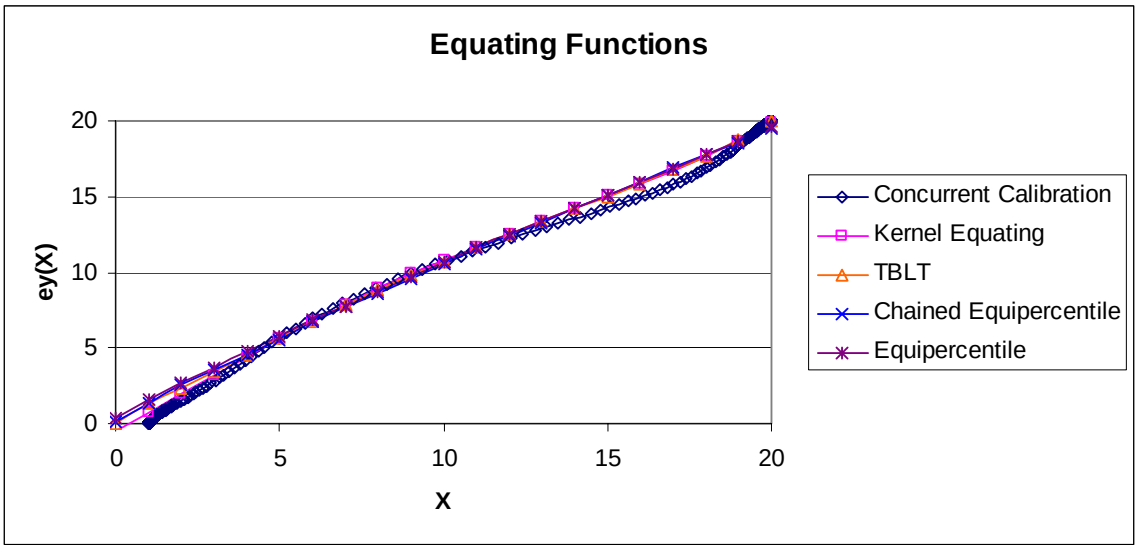


Figure E.80: 20 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

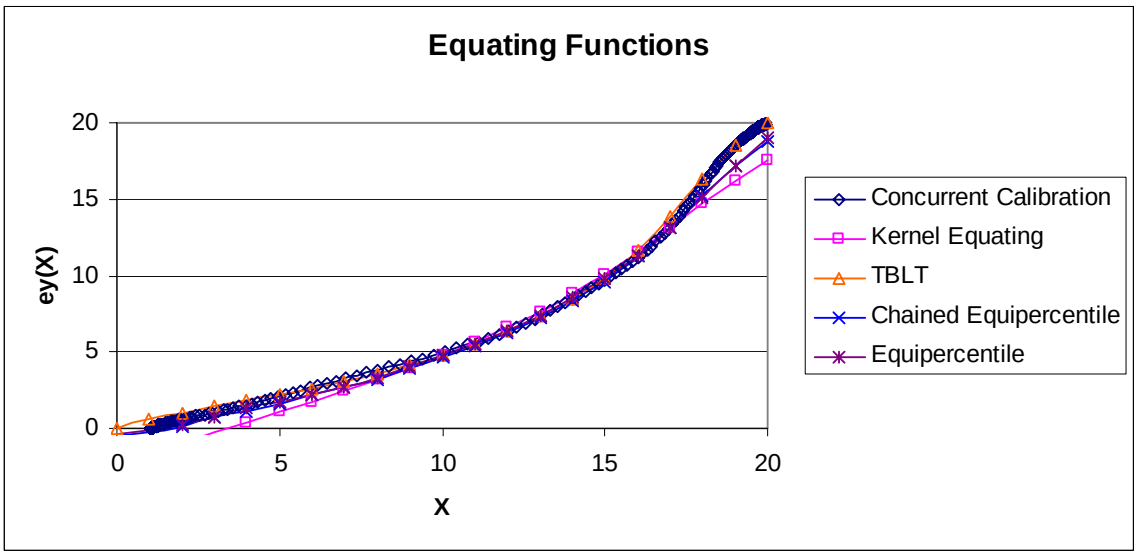


Figure E.81: 20 Items per form, 50% Anchor Length, 100,000 Sample Size, No Ability Difference

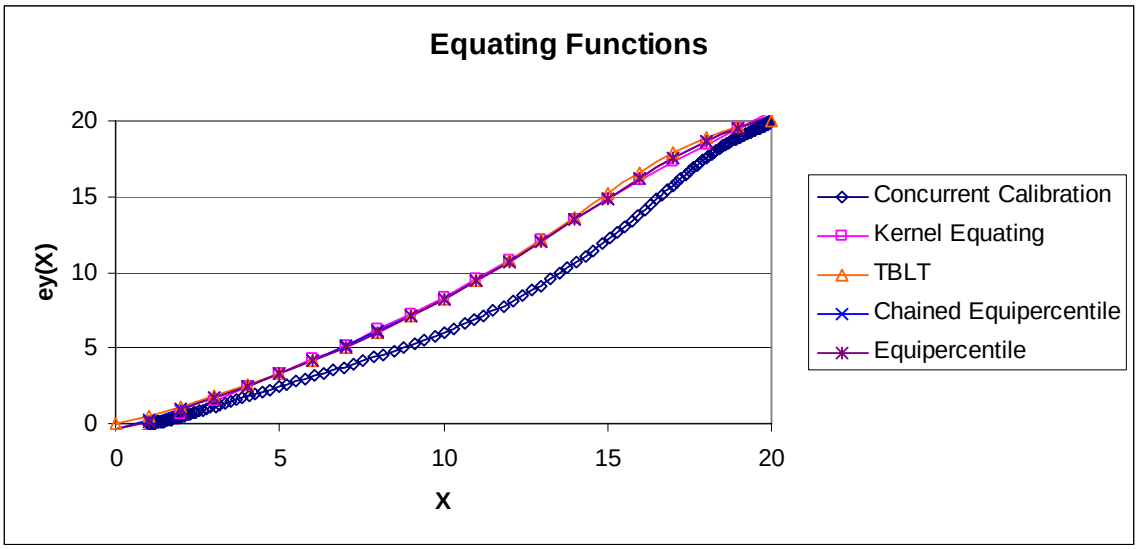


Figure E.82: 20 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

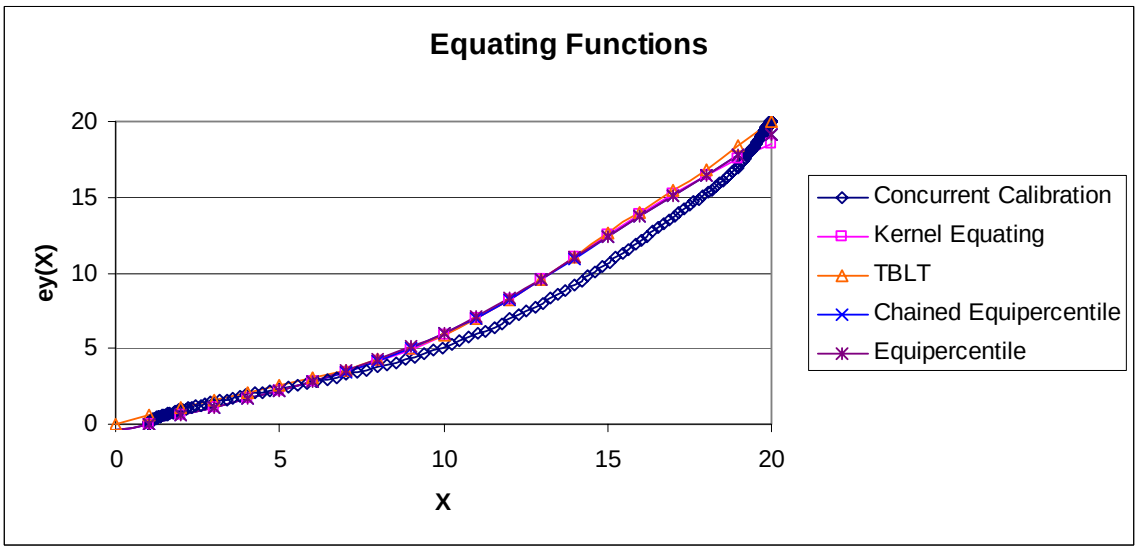


Figure E.83: 20 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

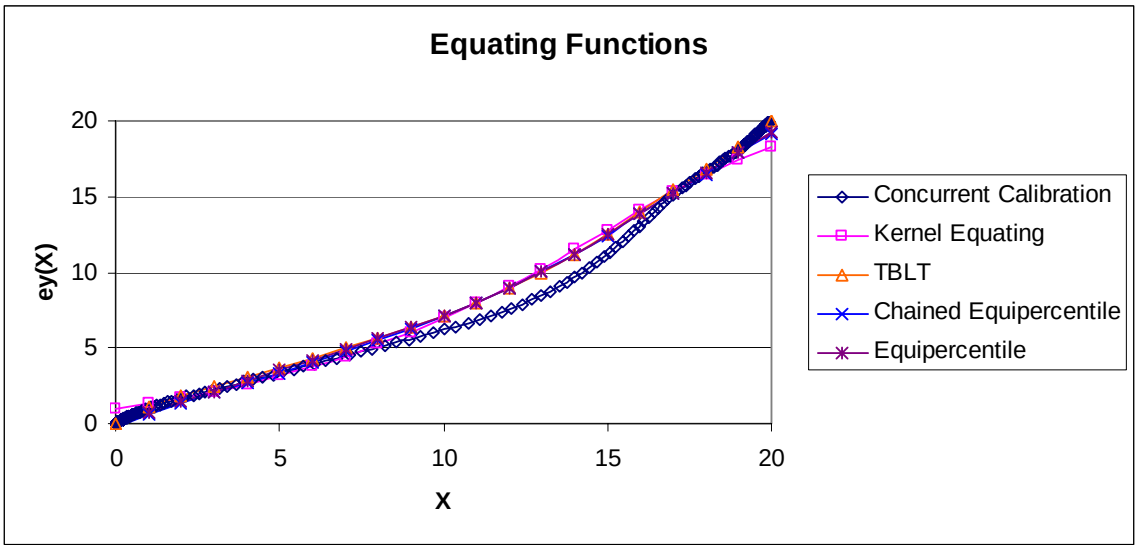
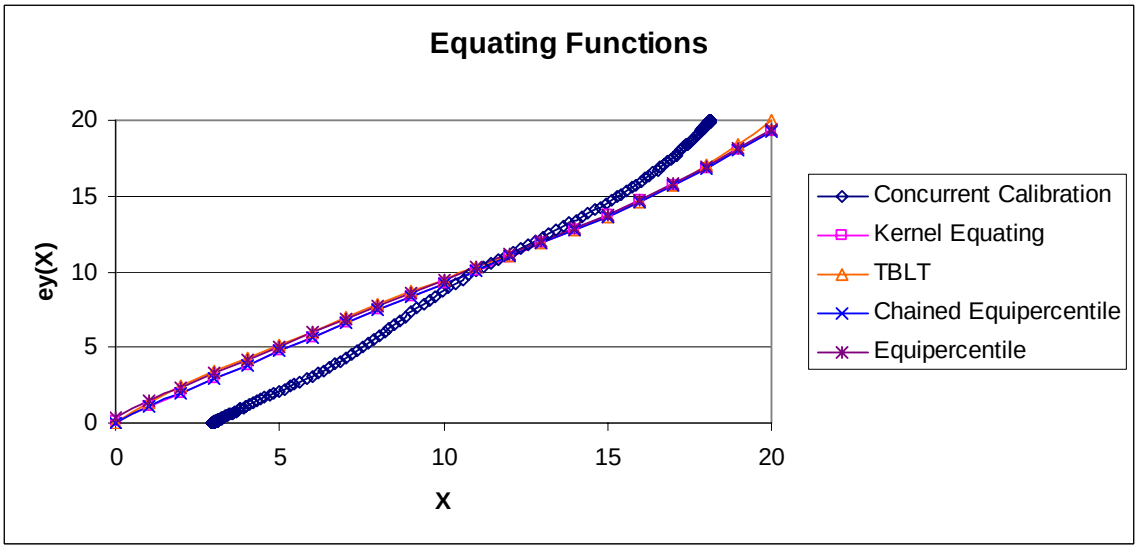


Figure E.84: 20 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference



APPENDIX F: Equating Function Differences

Figure F.1 100 Items per form, 50% Anchor Length, 1000 Sample Size, No Theta Difference

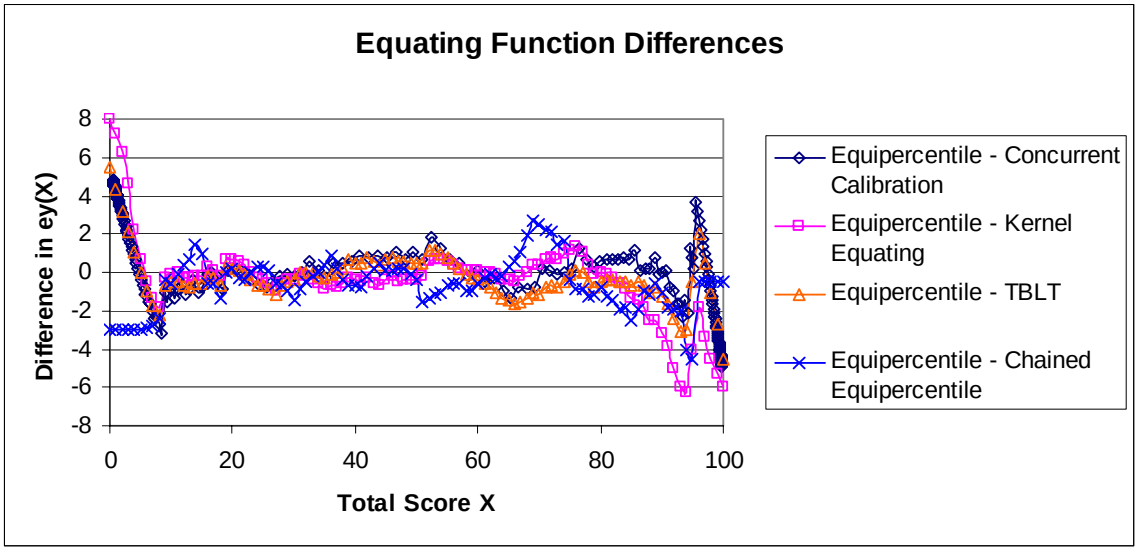


Figure F.2 100 Items per form, 50% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

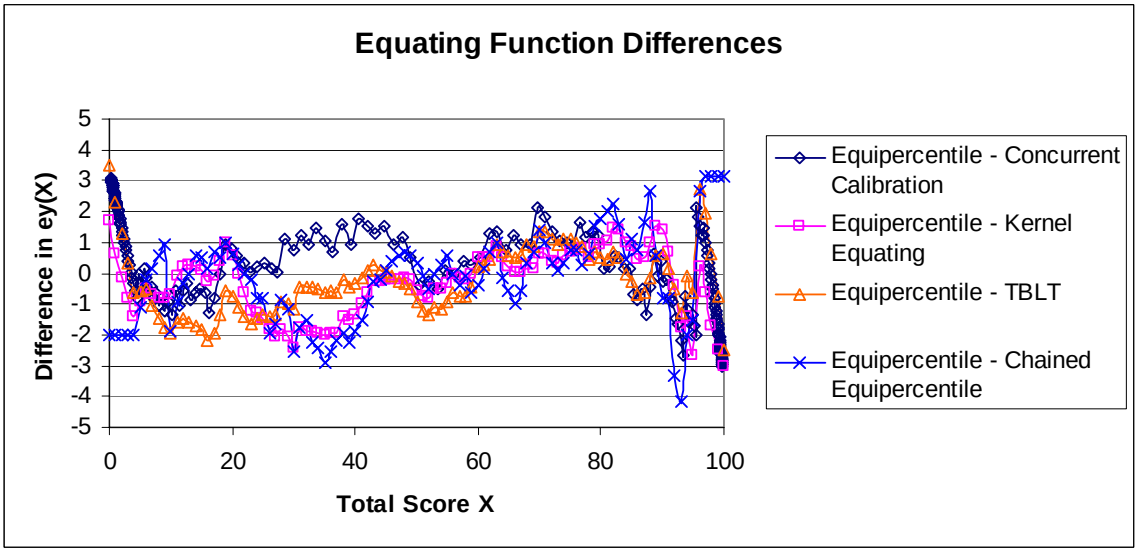


Figure F.3 100 Items per form, 50% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

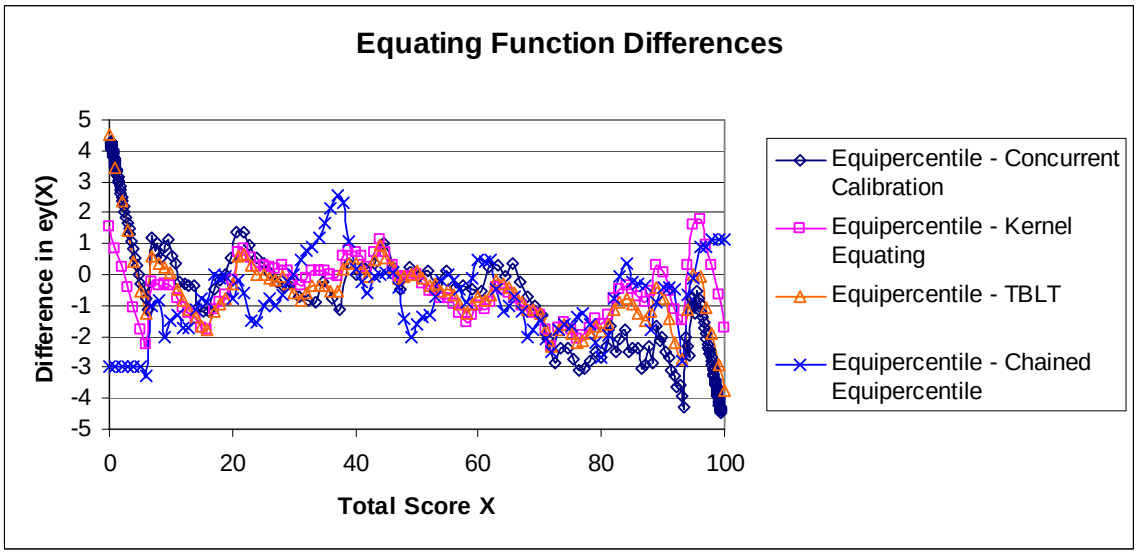


Figure F.4 100 Items per form, 50% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

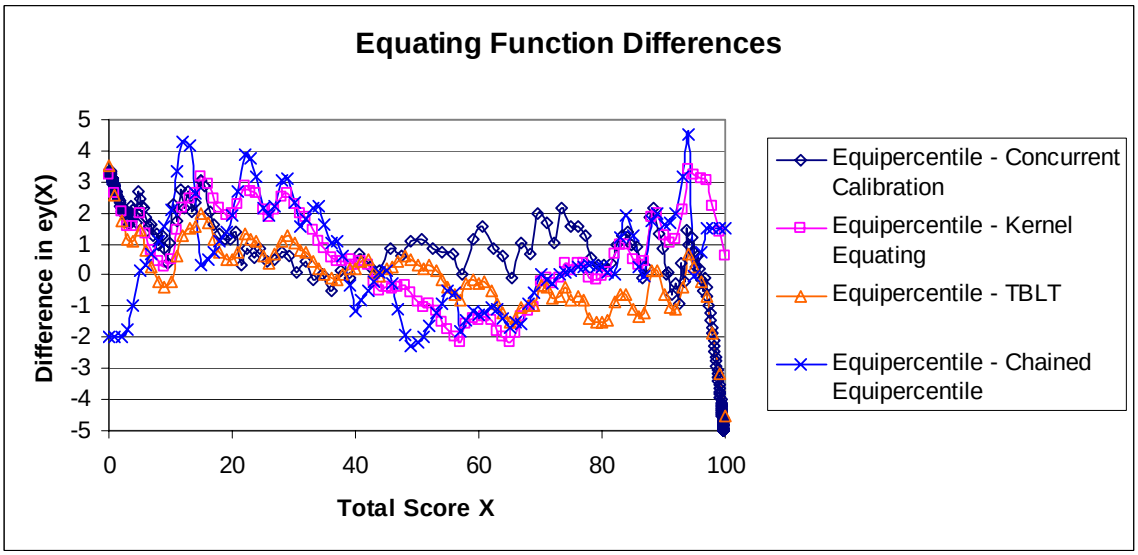


Figure F.5 100 Items per form, 50% Anchor Length, 10,000 Sample Size, No Theta Difference

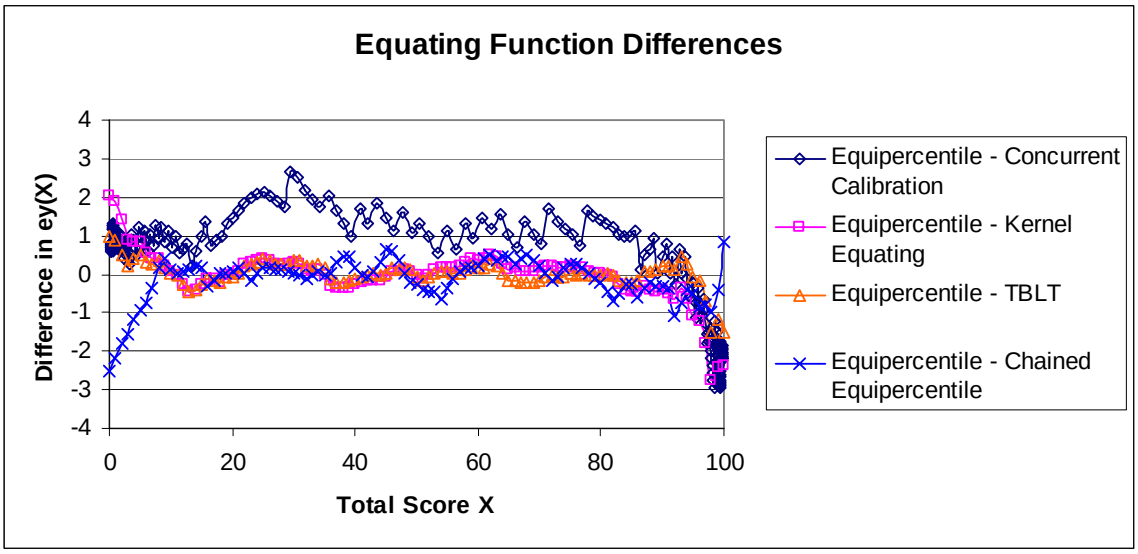


Figure F.6 100 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

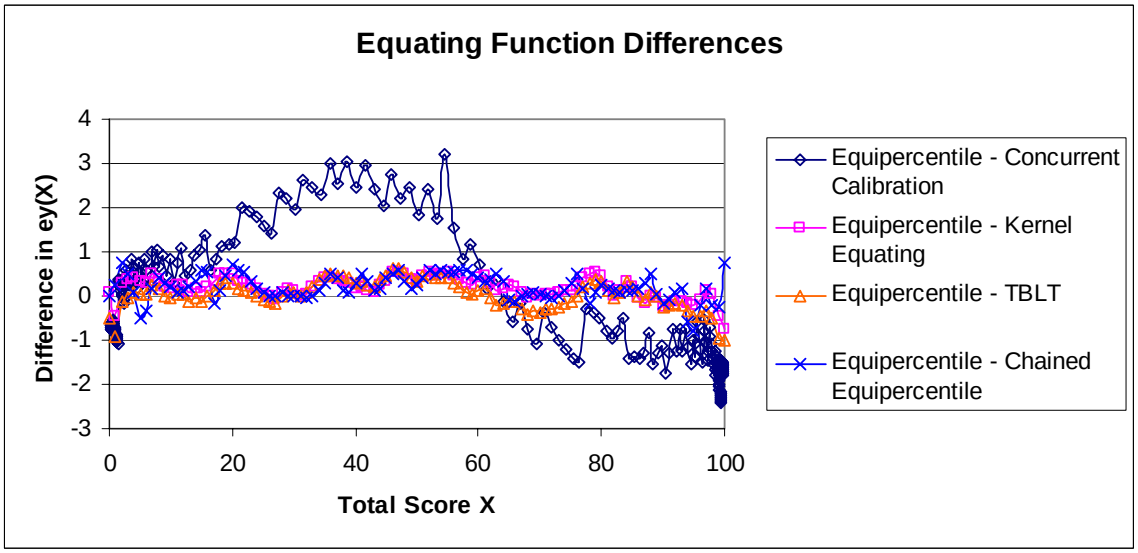


Figure F.7 100 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

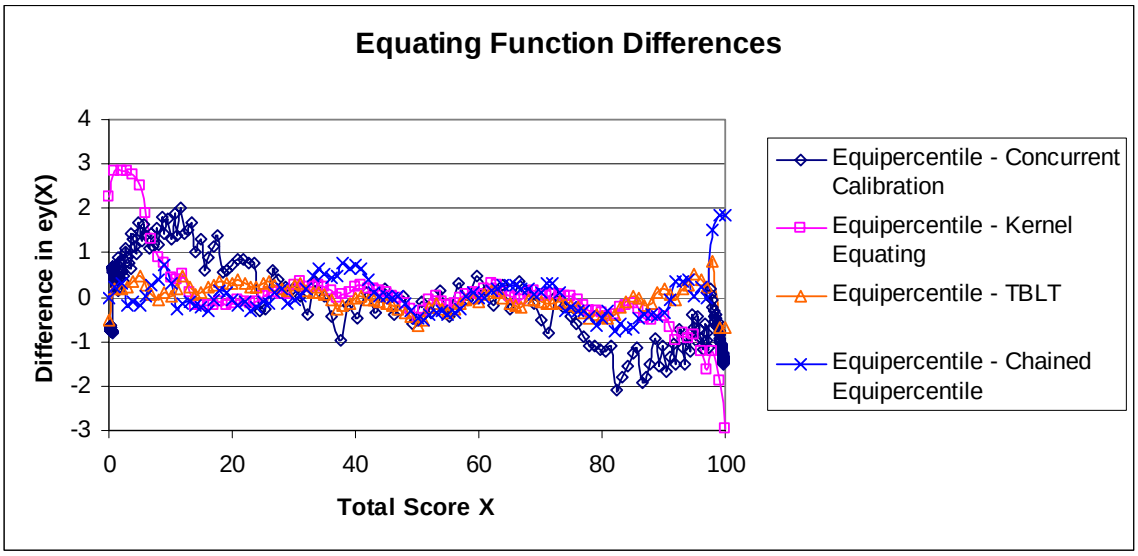


Figure F.8 100 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

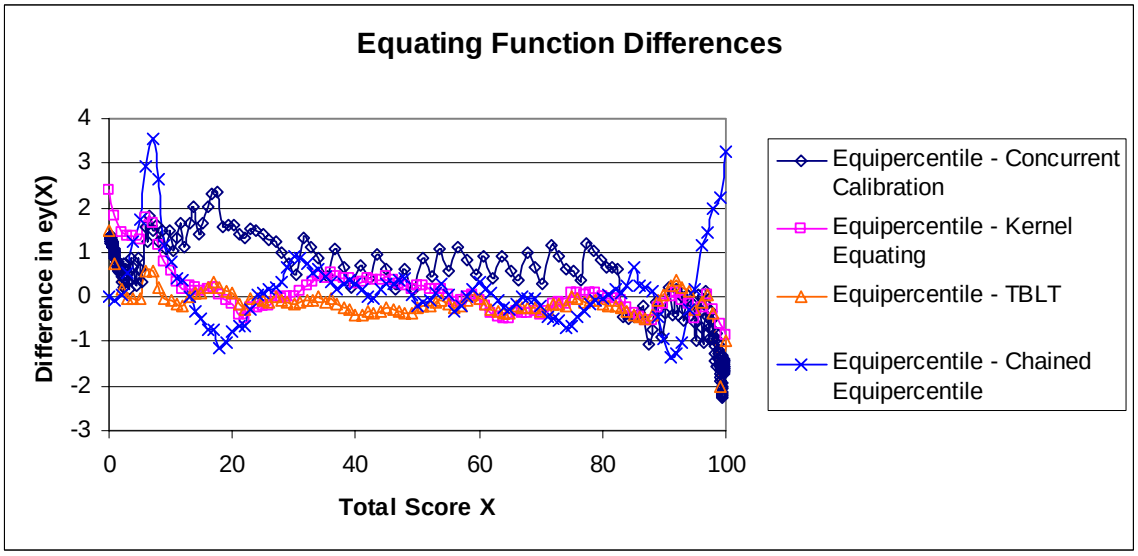


Figure F.9 100 Items per form, 50% Anchor Length, 100,000 Sample Size, No Theta Difference

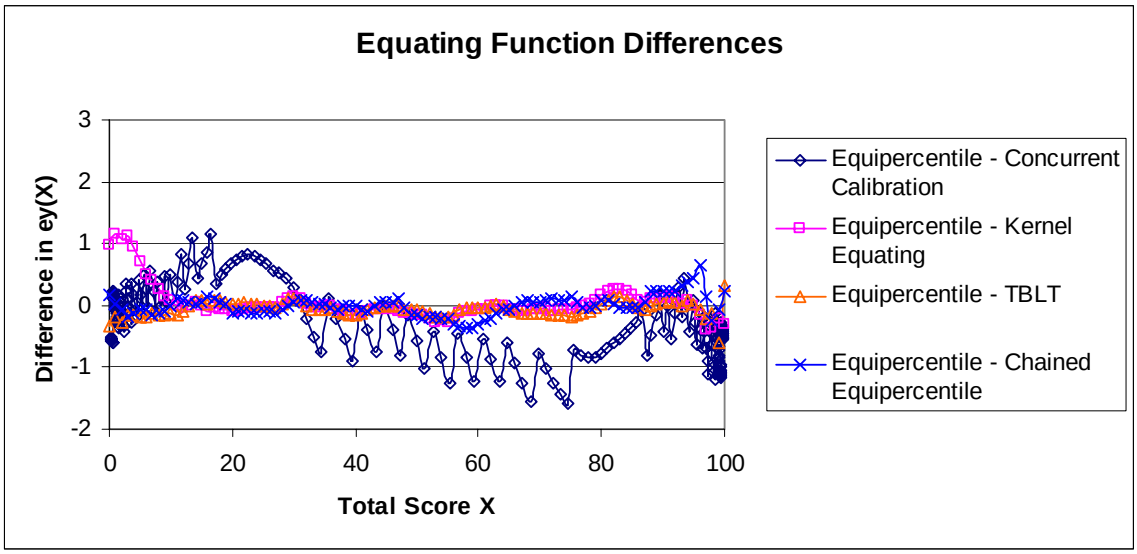


Figure F.10 100 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

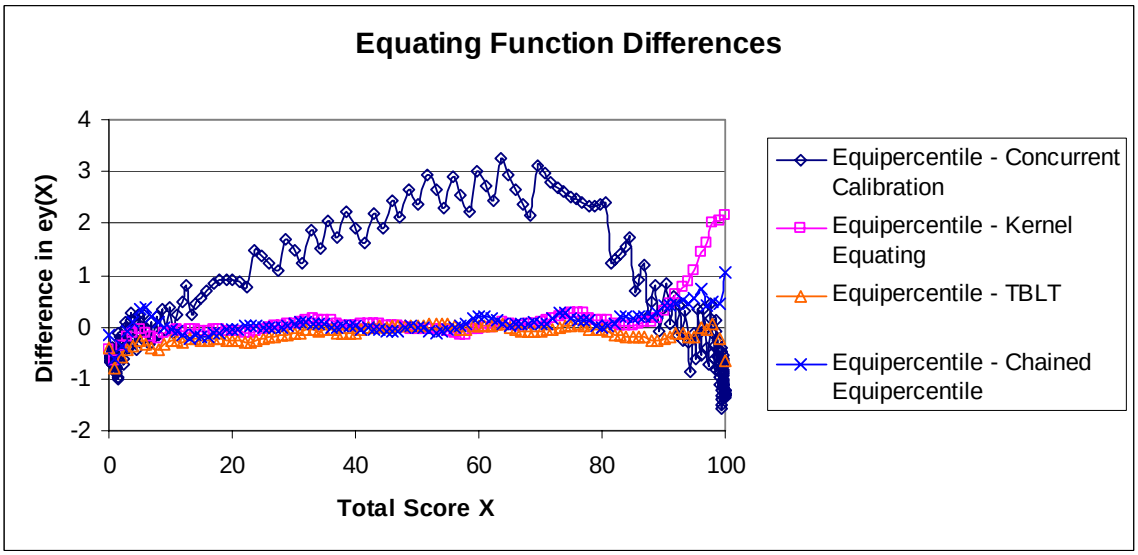


Figure F.11 100 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

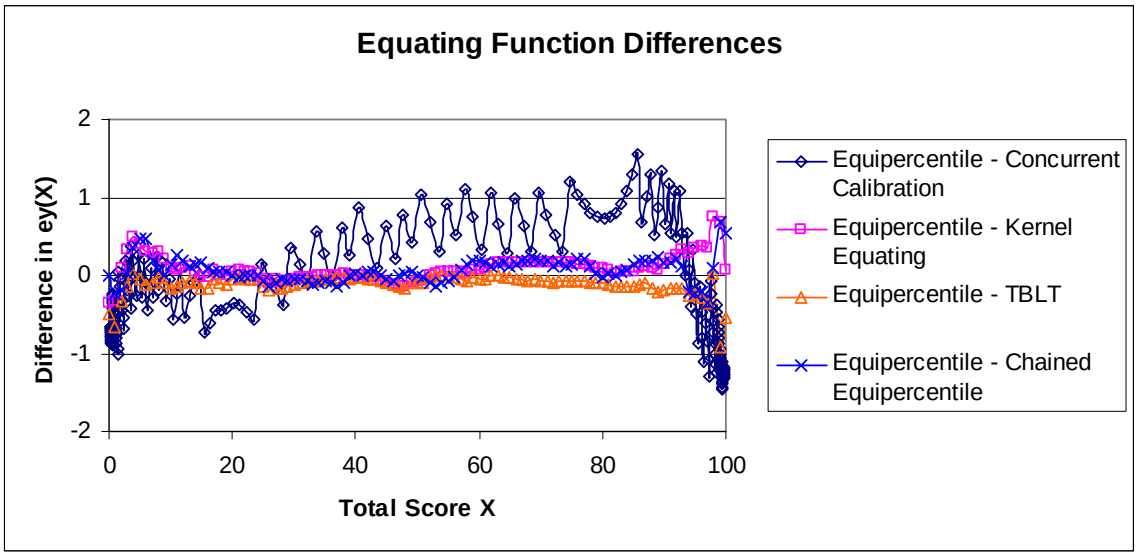


Figure F.12 100 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

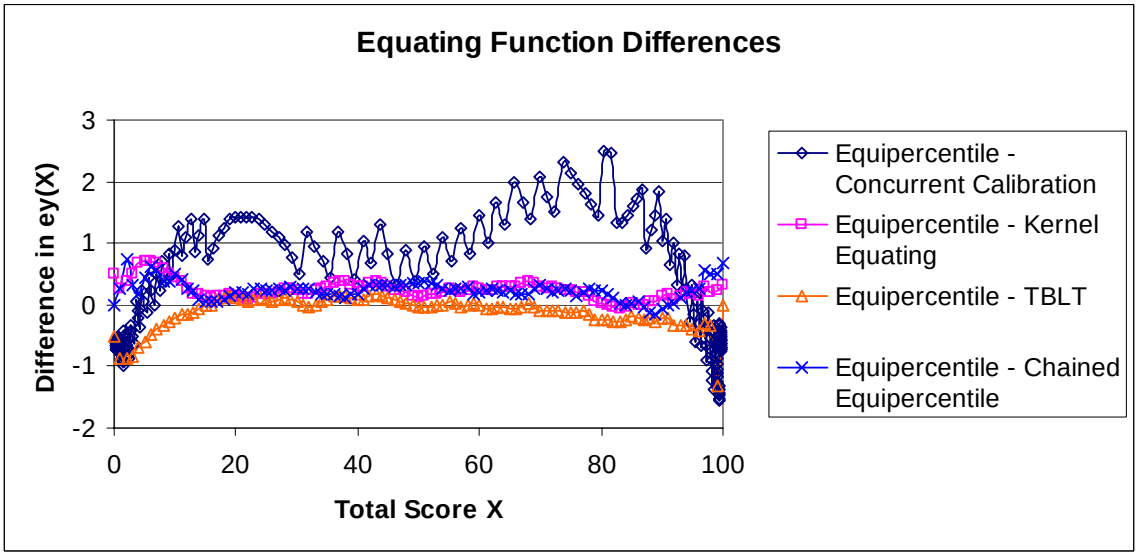


Figure F.13 100 Items per form, 35% Anchor Length, 1000 Sample Size, No Theta Difference

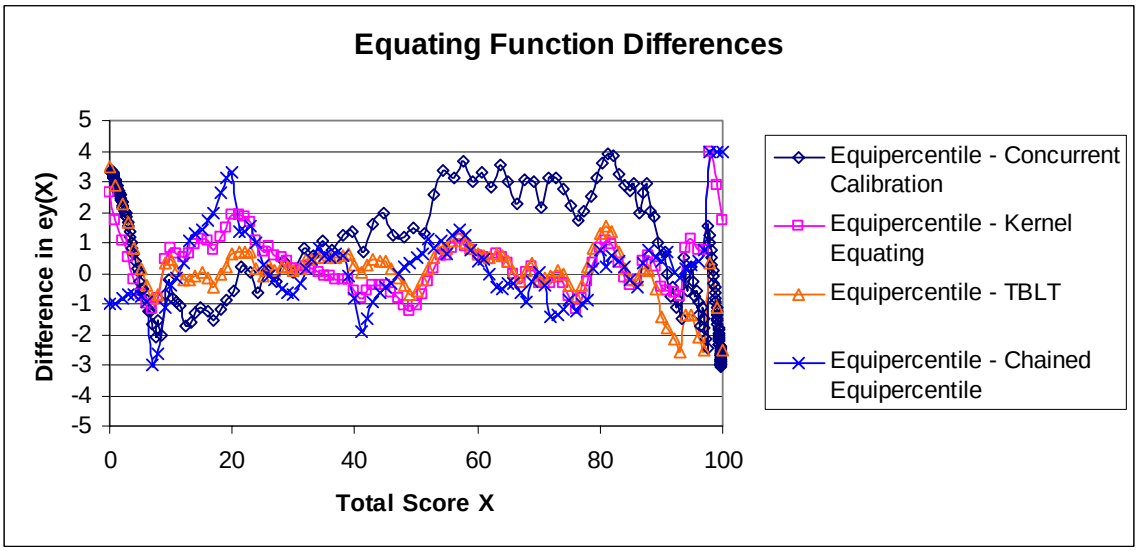


Figure F.14 100 Items per form, 35% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

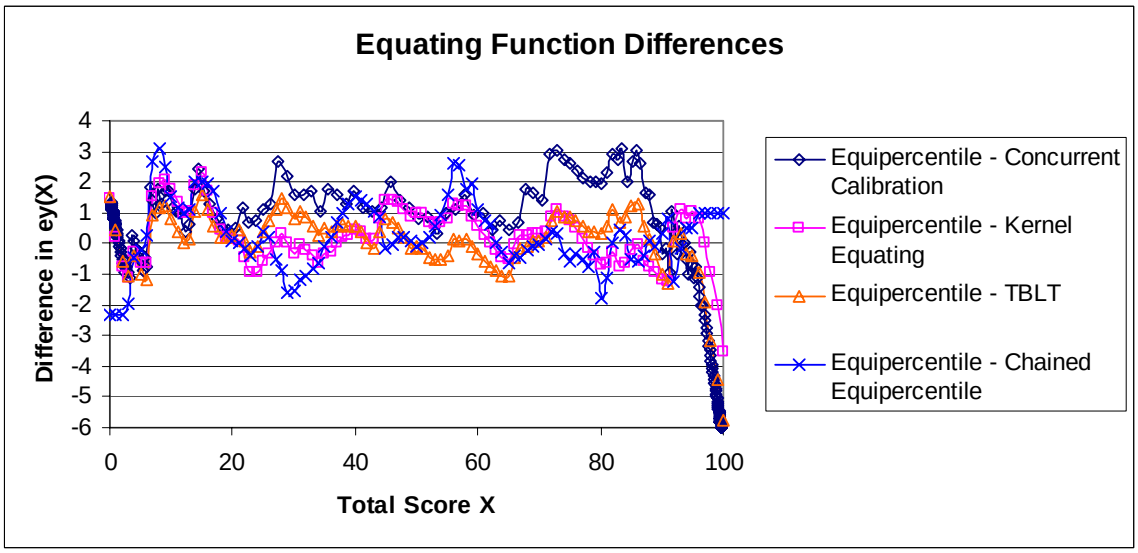


Figure F.15 100 Items per form, 35% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

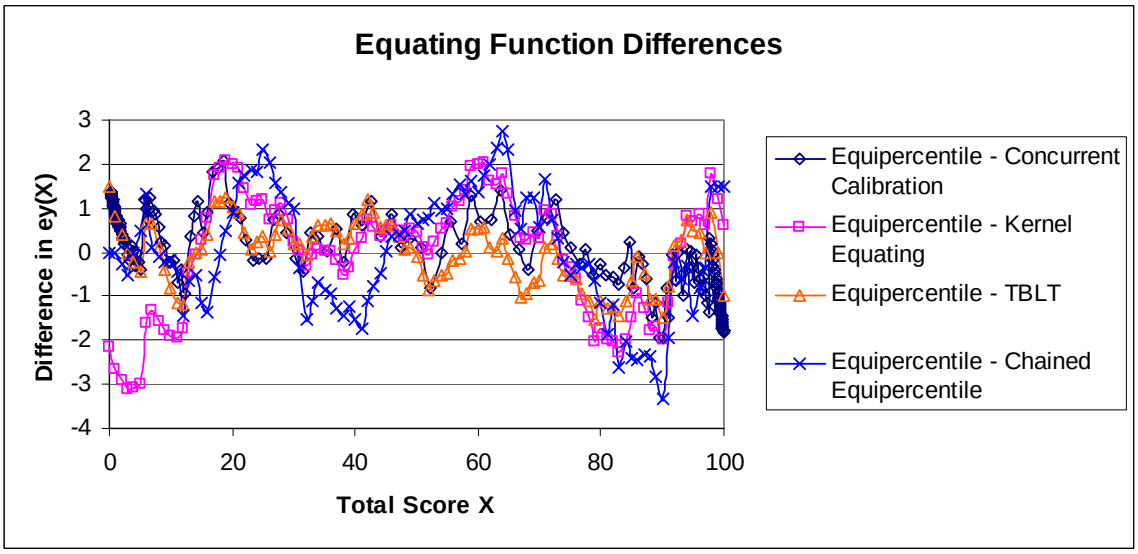


Figure F.16 100 Items per form, 35% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

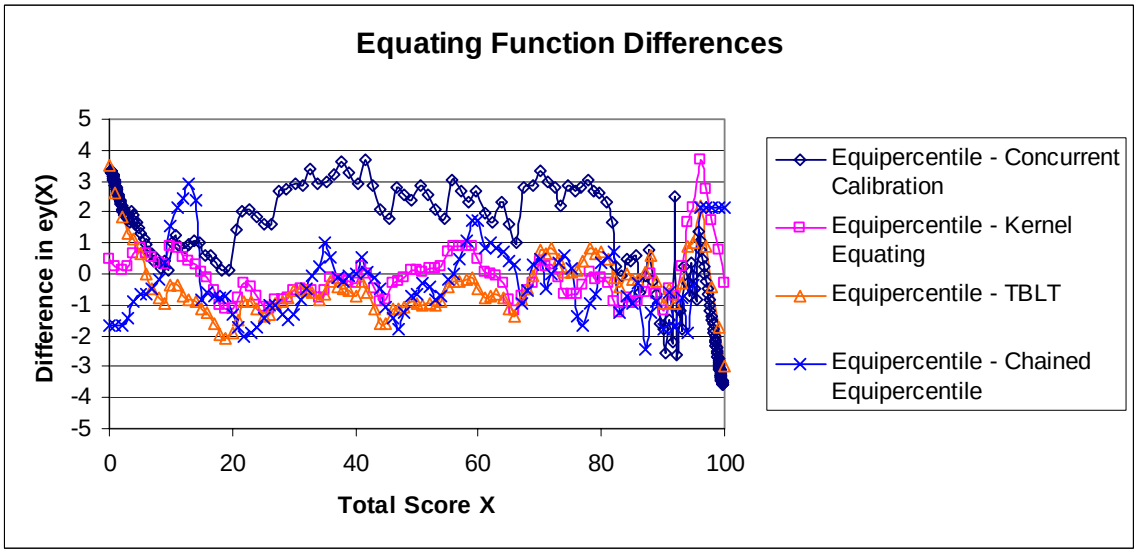


Figure F.17 100 Items per form, 35% Anchor Length, 10,000 Sample Size, No Theta Difference

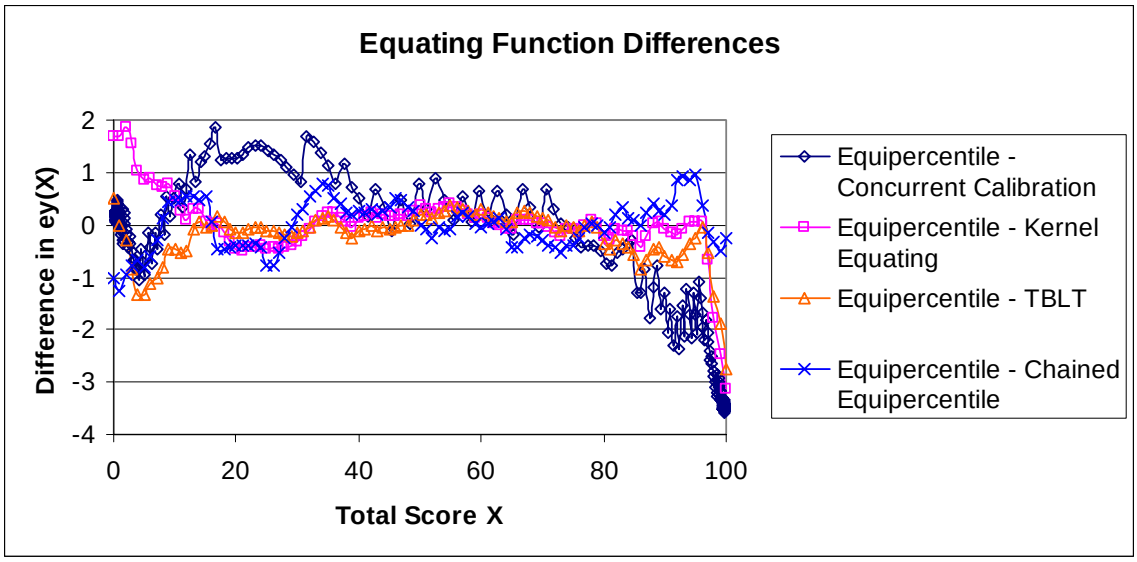


Figure F.18 100 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

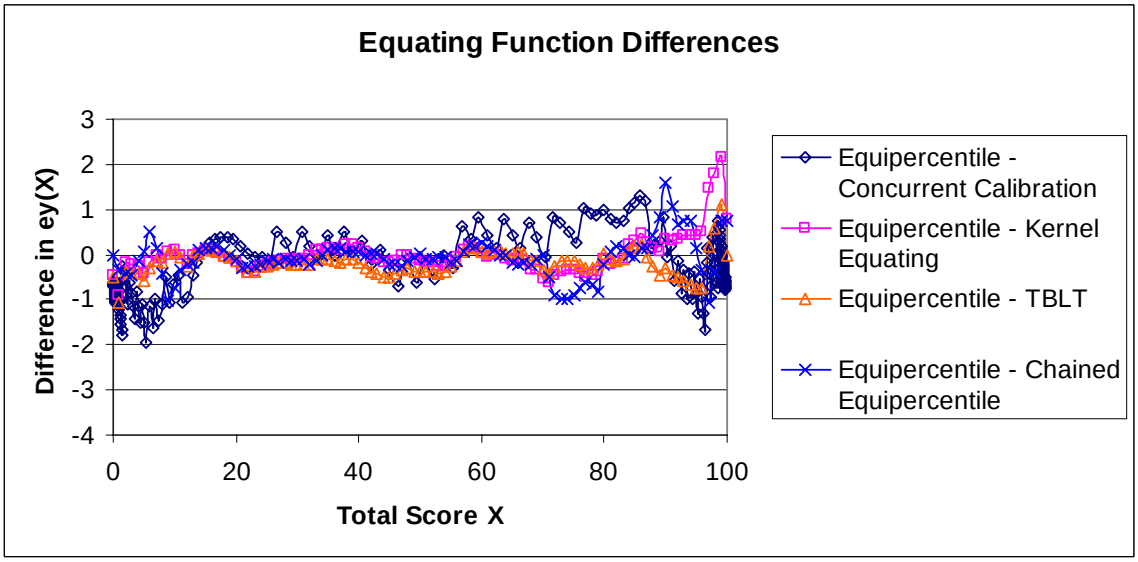


Figure F.19 100 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

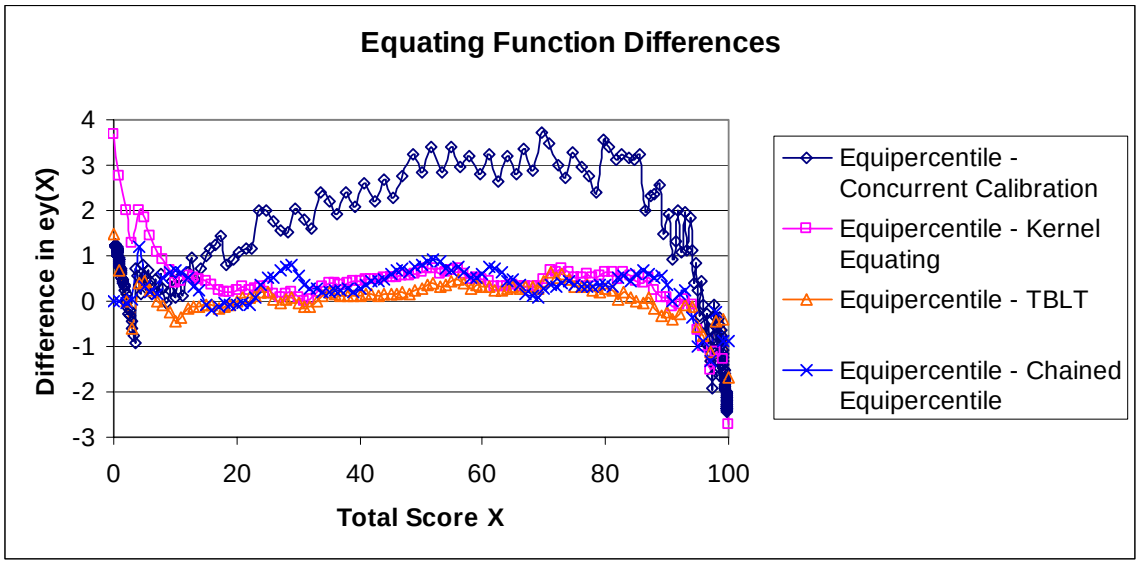


Figure F.20 100 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

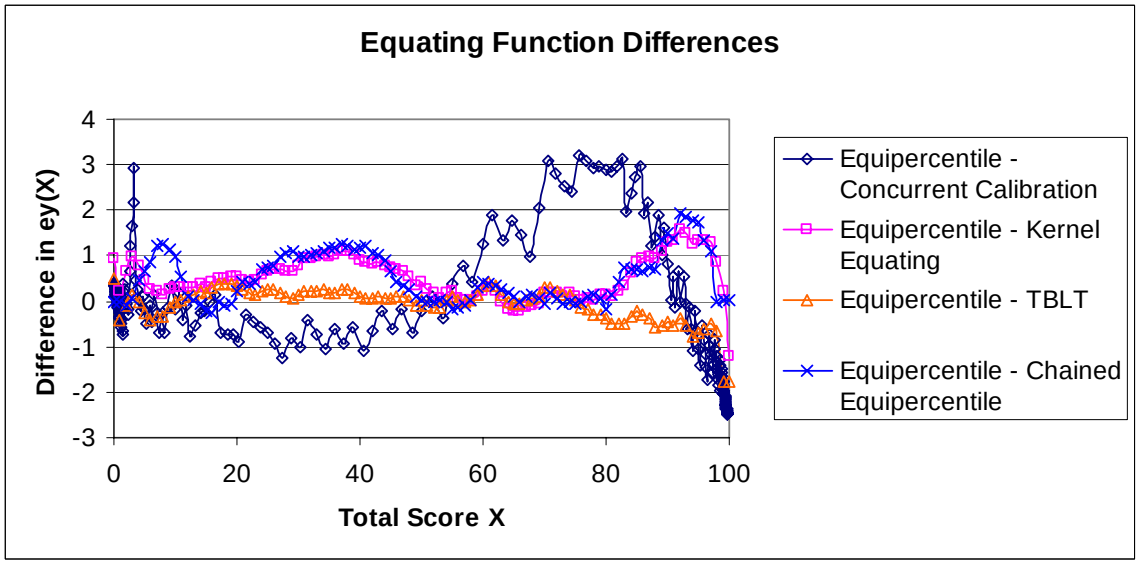


Figure F.21 100 Items per form, 35% Anchor Length, 100,000 Sample Size, No Theta Difference

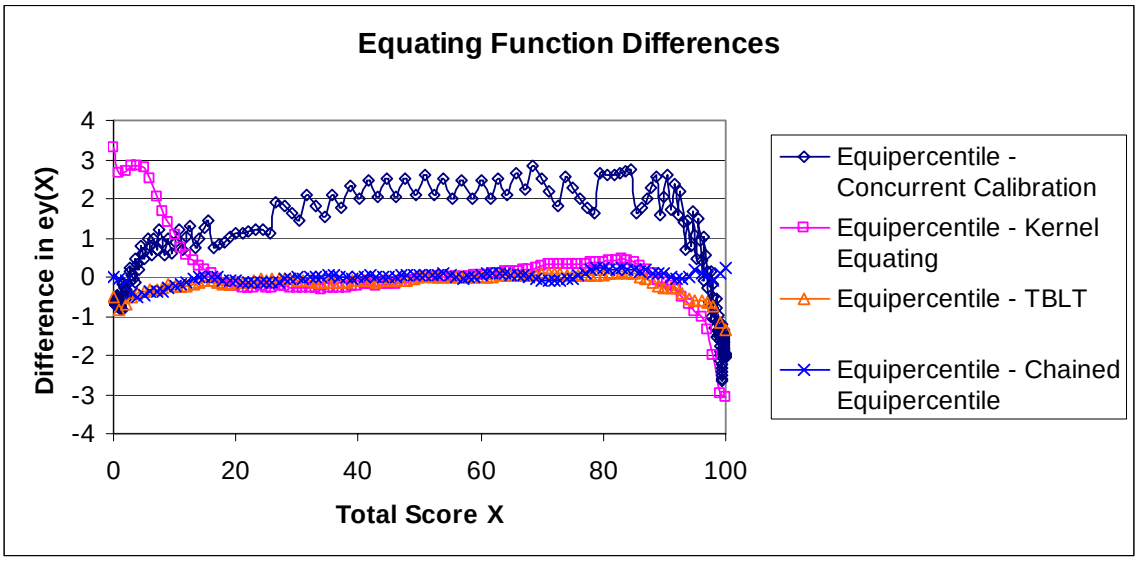


Figure F.22 100 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

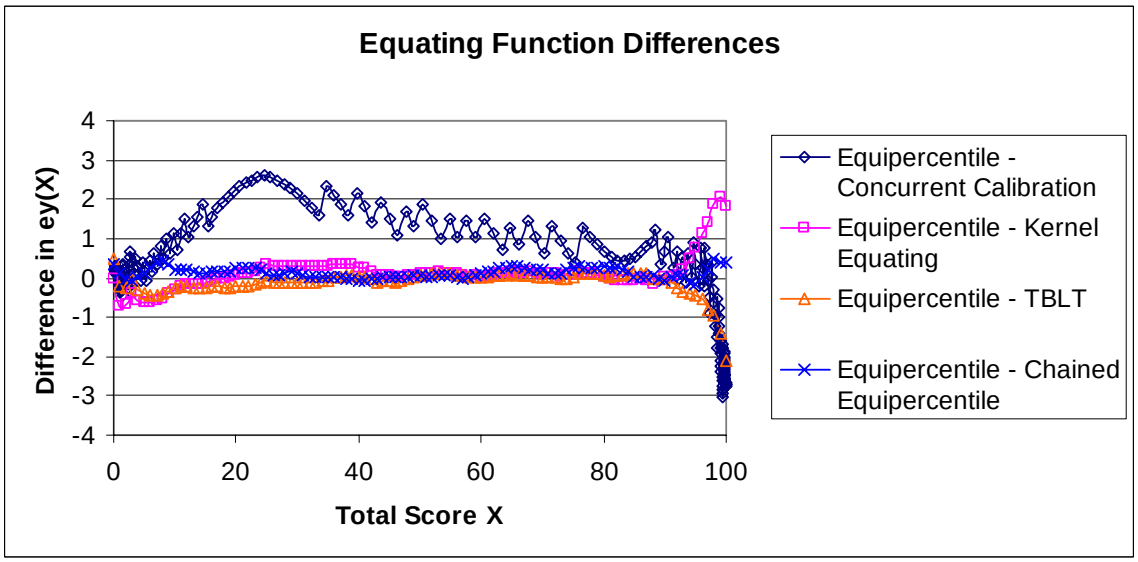


Figure F.23 100 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

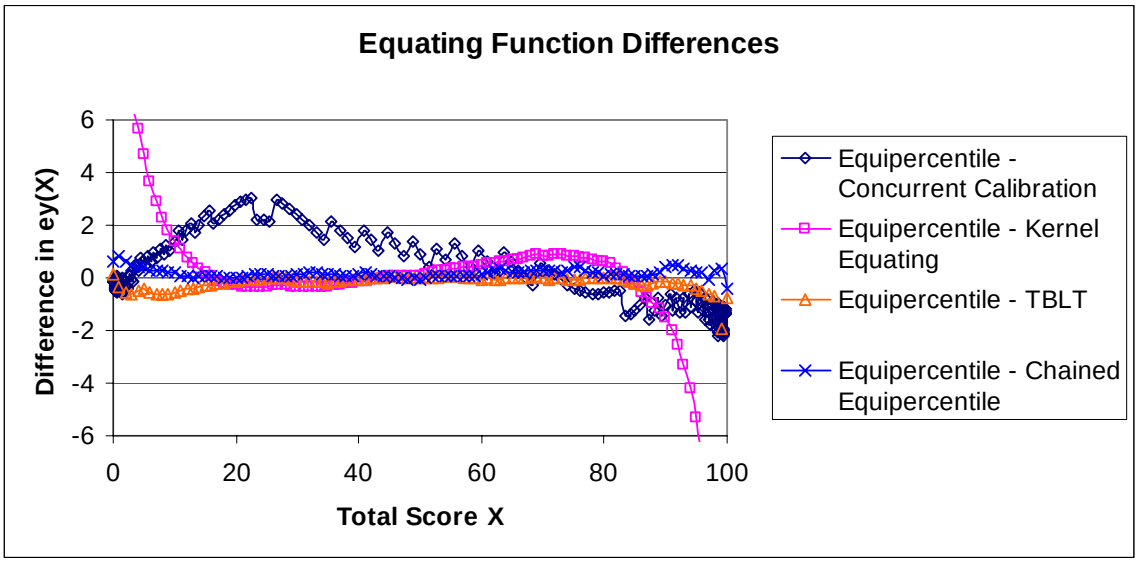


Figure F.24 100 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

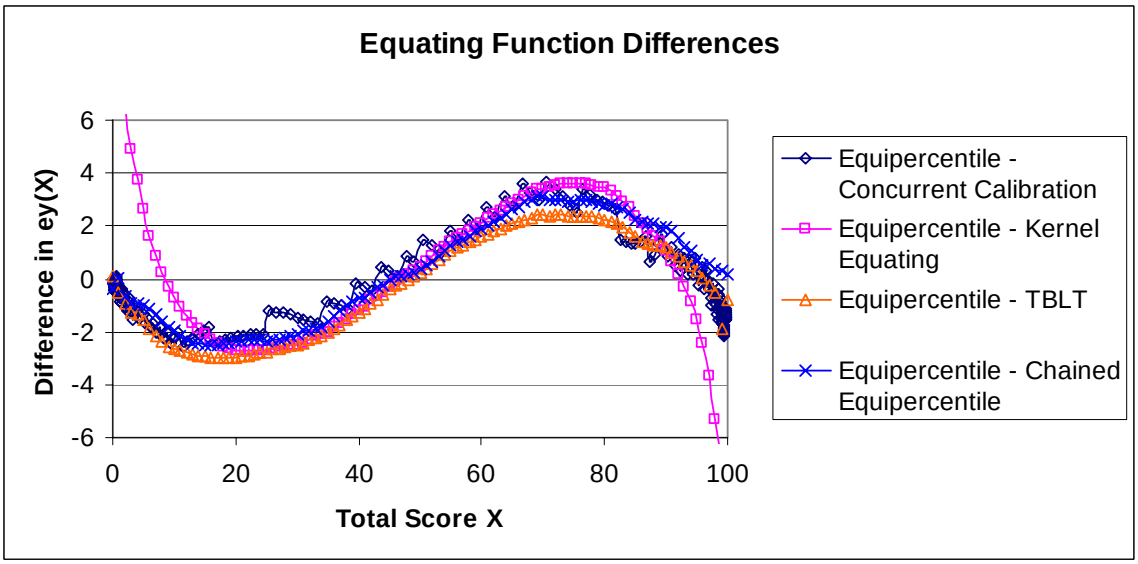


Figure F.25 100 Items per form, 20% Anchor Length, 1000 Sample Size, No Theta Difference

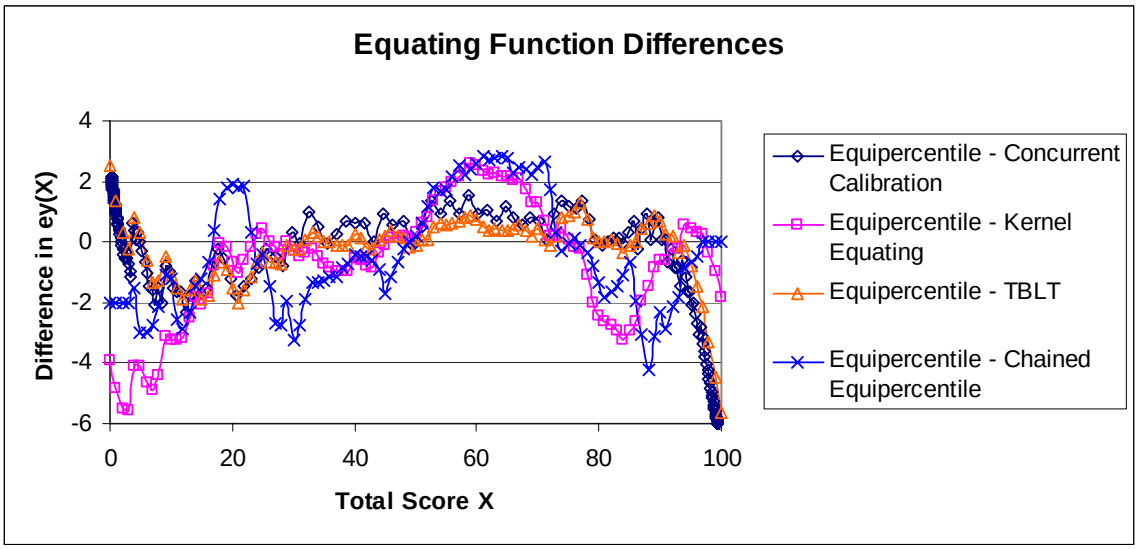


Figure F.26 100 Items per form, 20% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

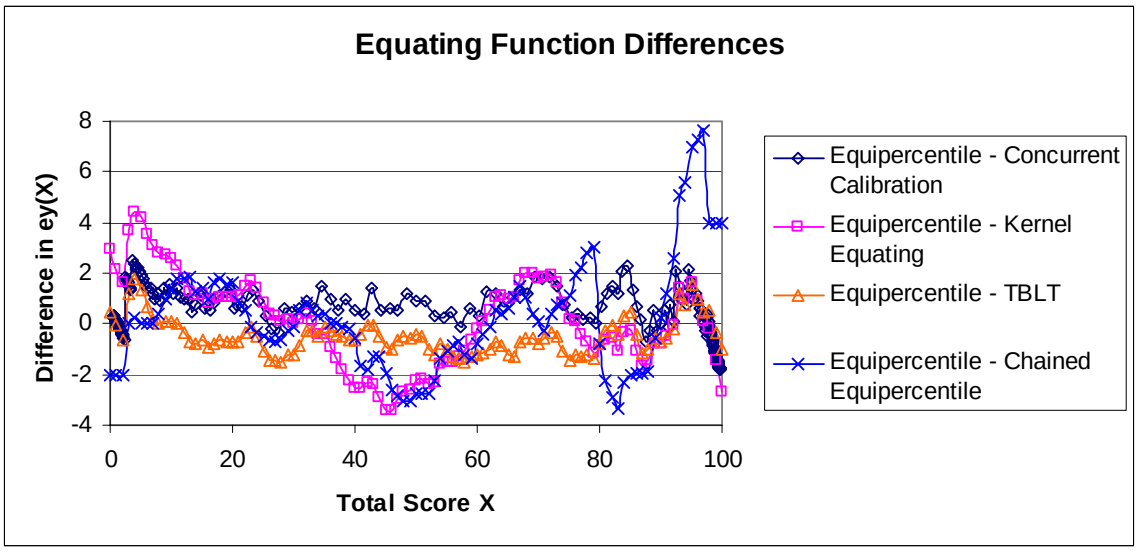


Figure F.27 100 Items per form, 20% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

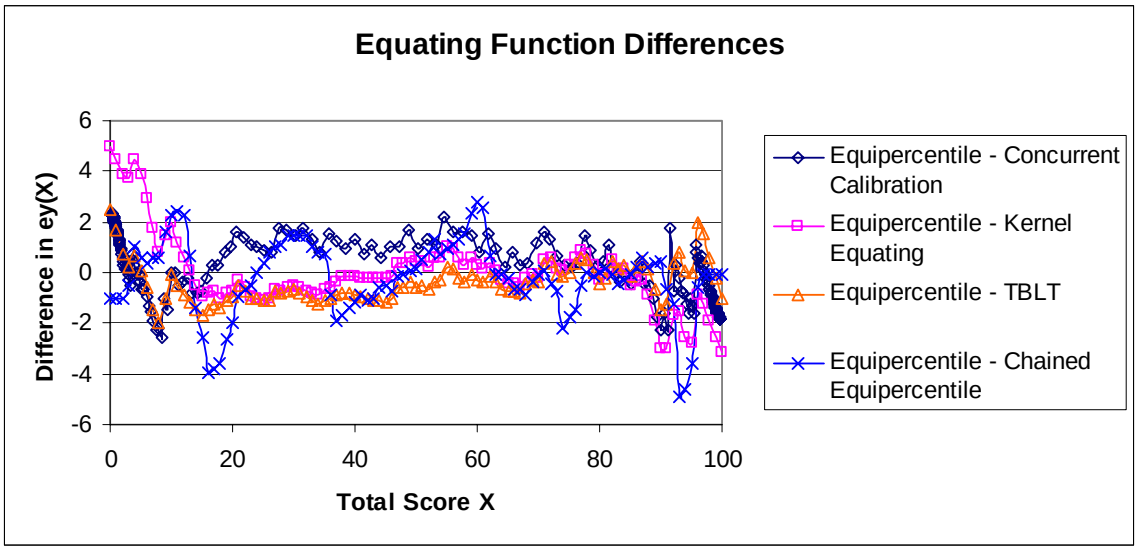


Figure F.28 100 Items per form, 20% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

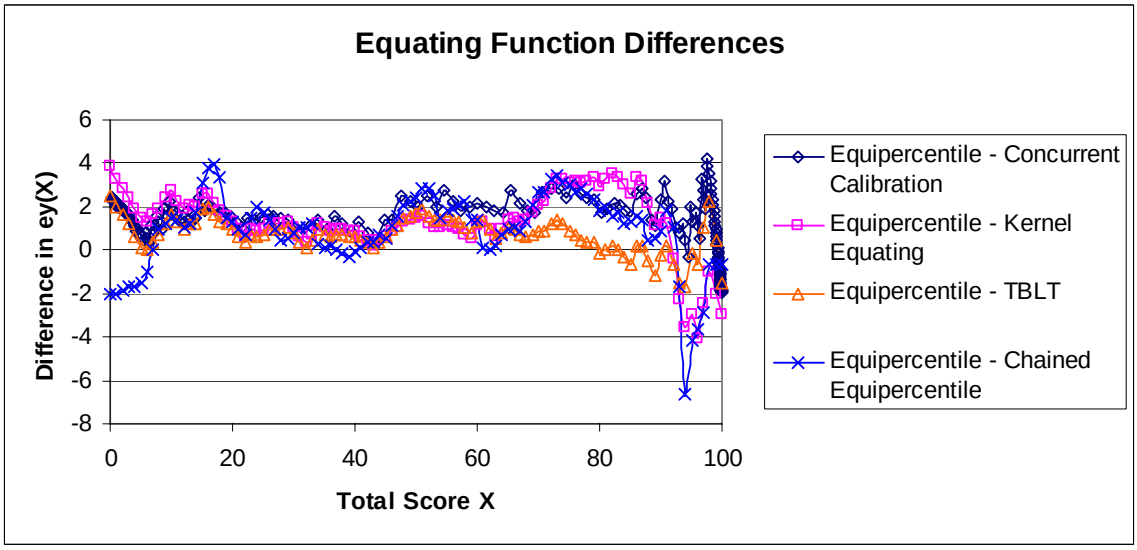


Figure F.29 100 Items per form, 20% Anchor Length, 10,000 Sample Size, No Theta Difference

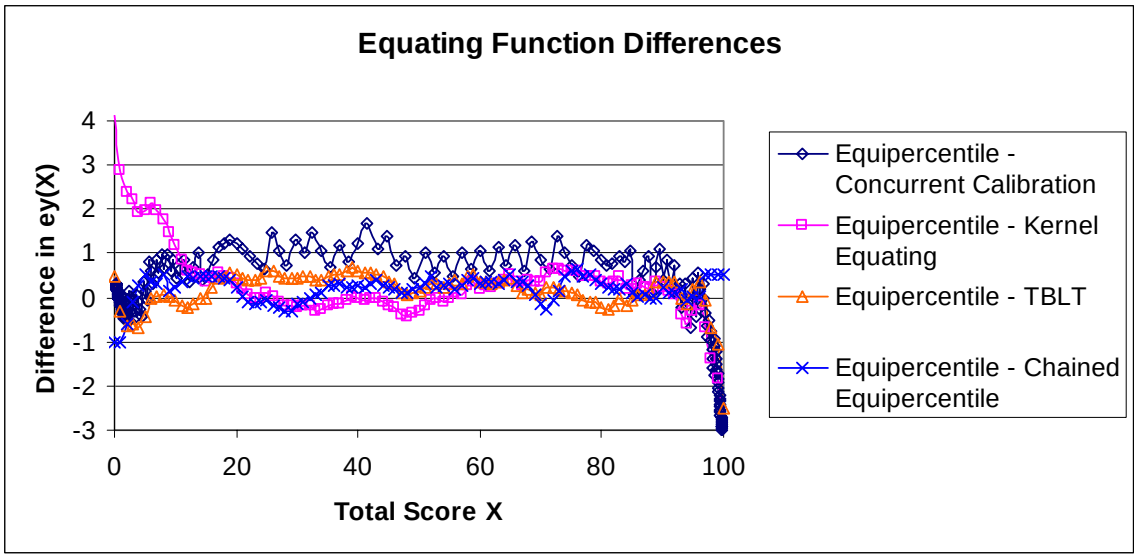


Figure F.30 100 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

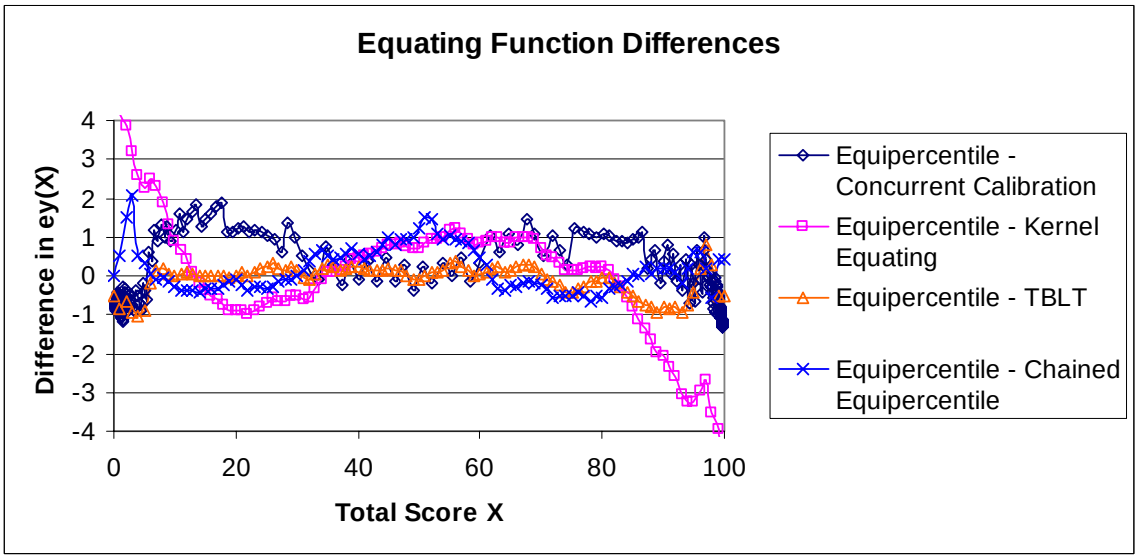


Figure F.31 100 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

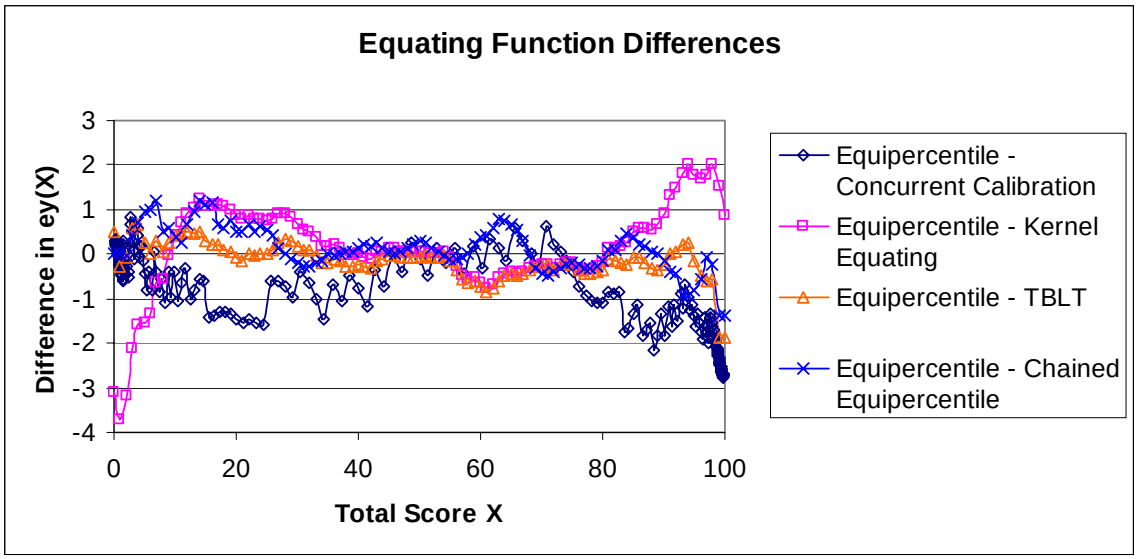


Figure F.32 100 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

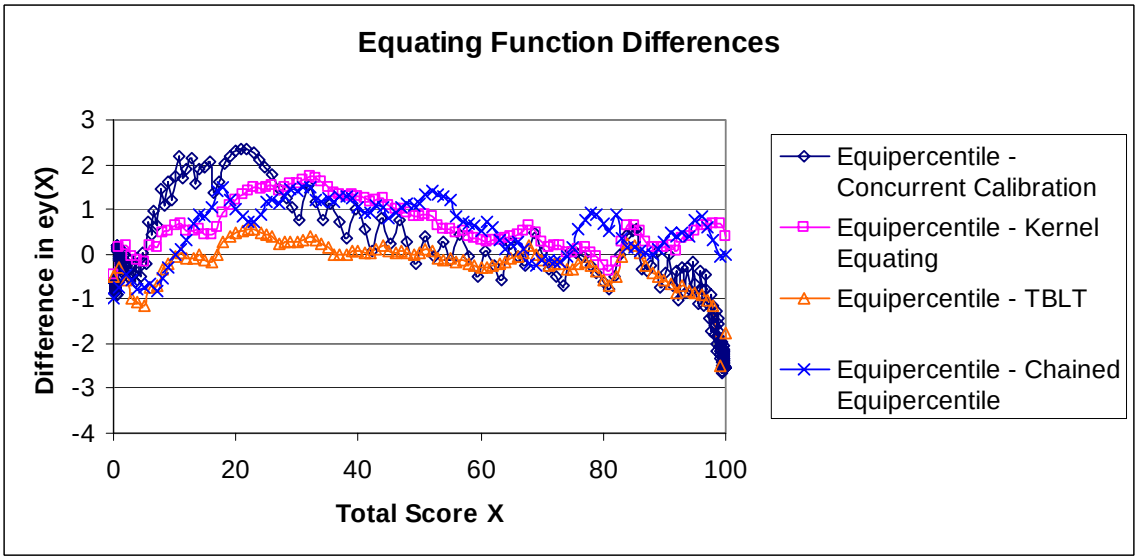


Figure F.33 100 Items per form, 20% Anchor Length, 100,000 Sample Size, No Theta Difference

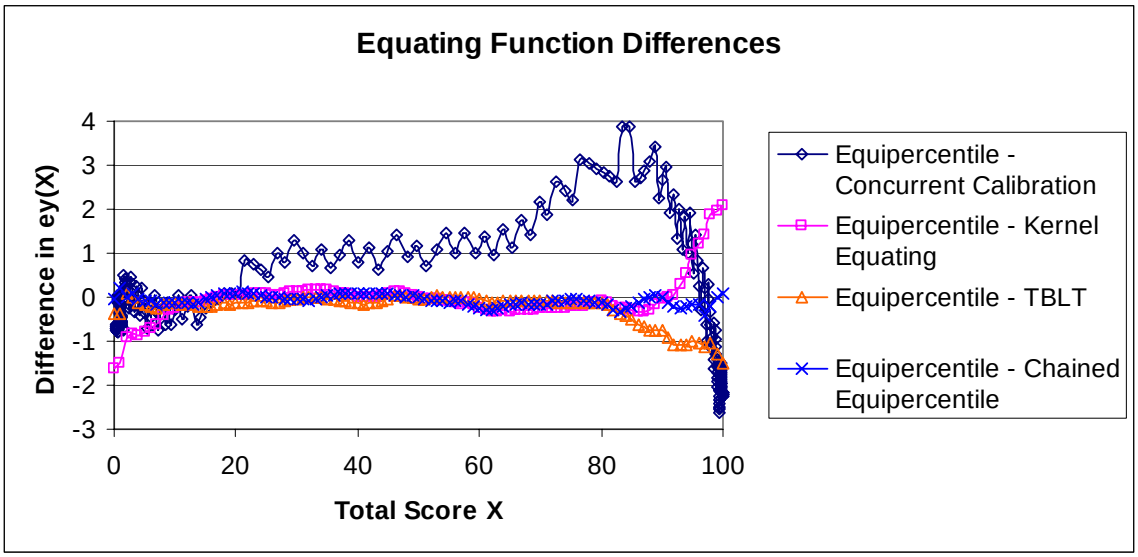


Figure F.34 100 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

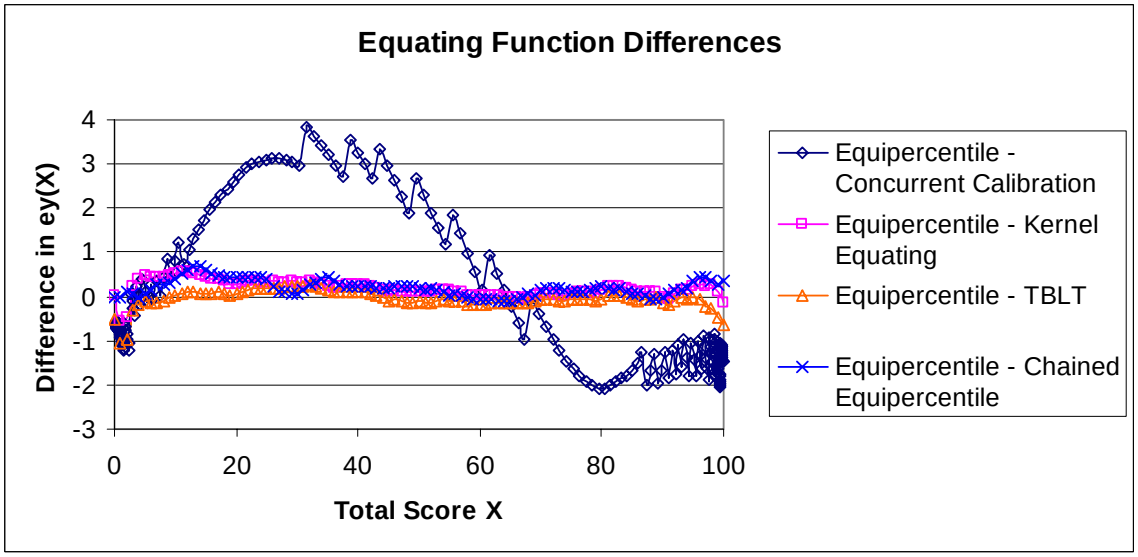


Figure F.35 100 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

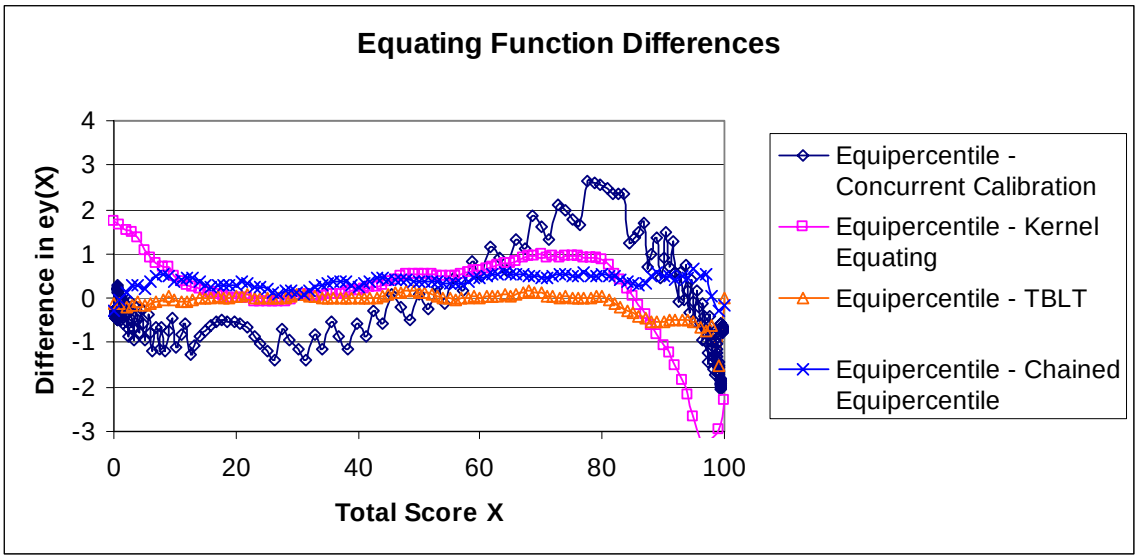


Figure F.36 100 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

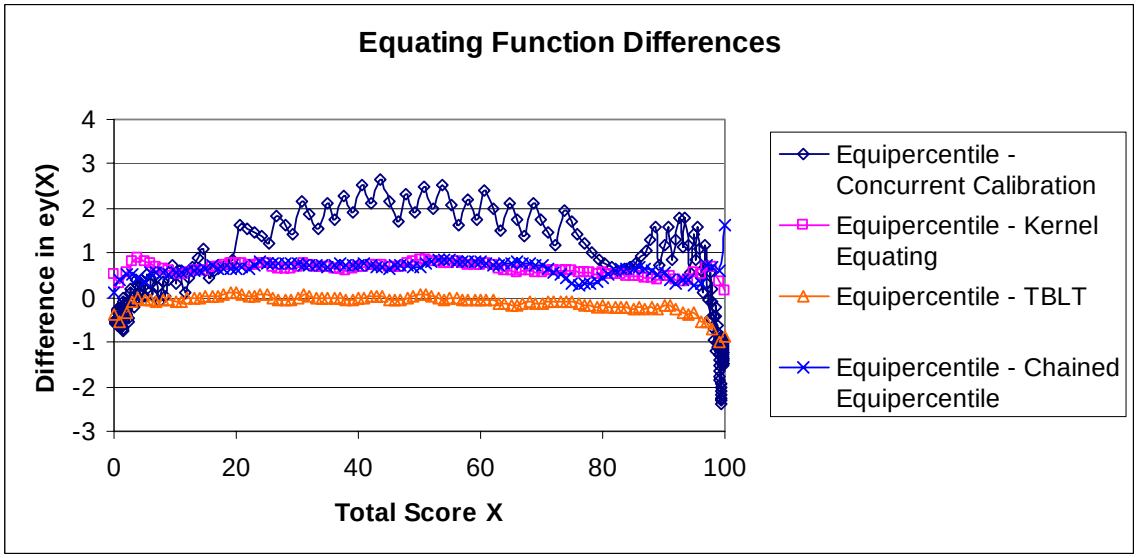


Figure F.37 60 Items per form, 50% Anchor Length, 1000 Sample Size, No Theta Difference

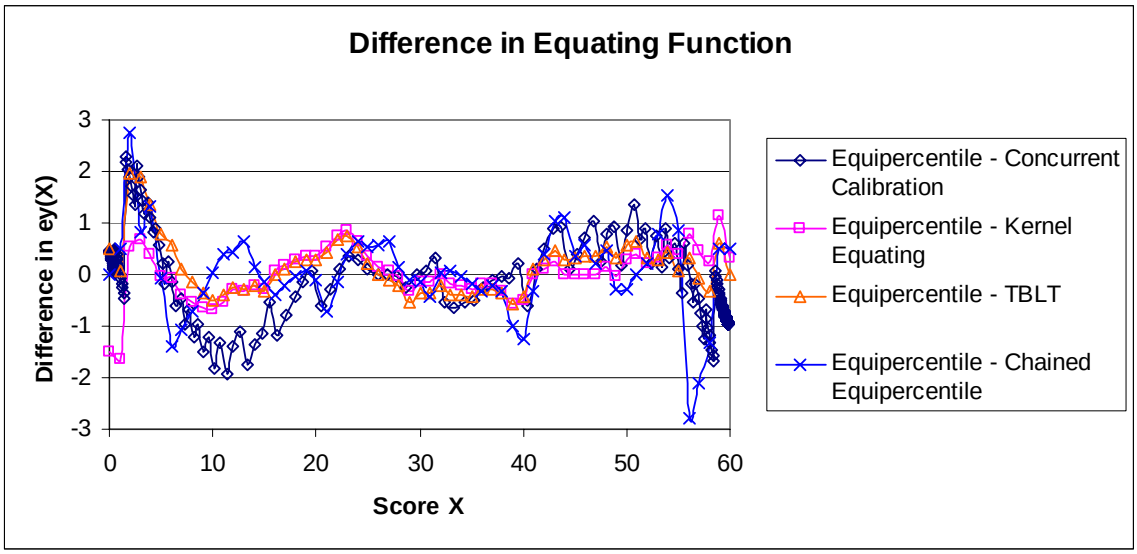


Figure F.38 60 Items per form, 50% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

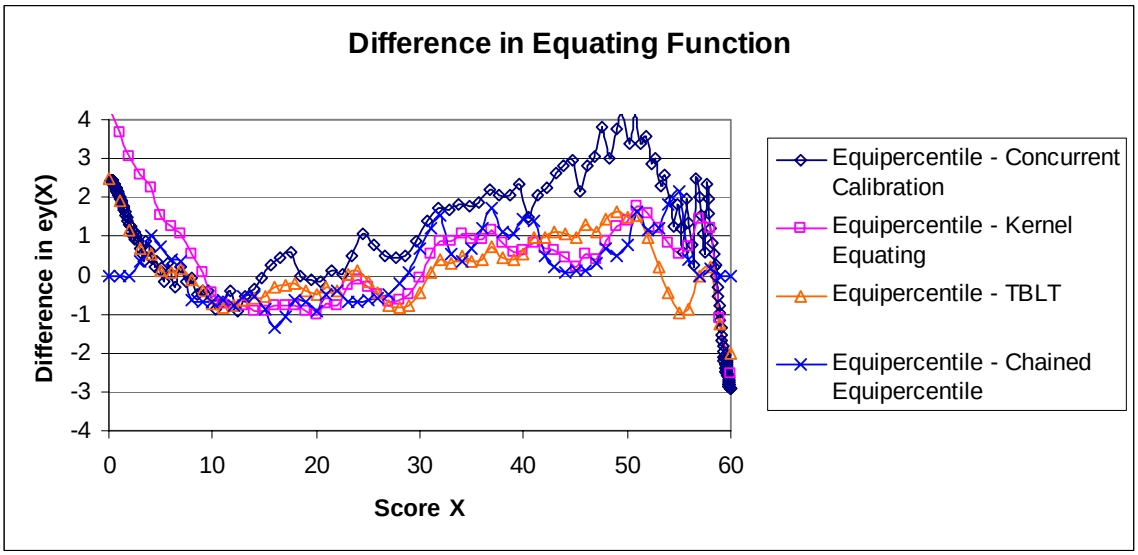


Figure F.39 60 Items per form, 50% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

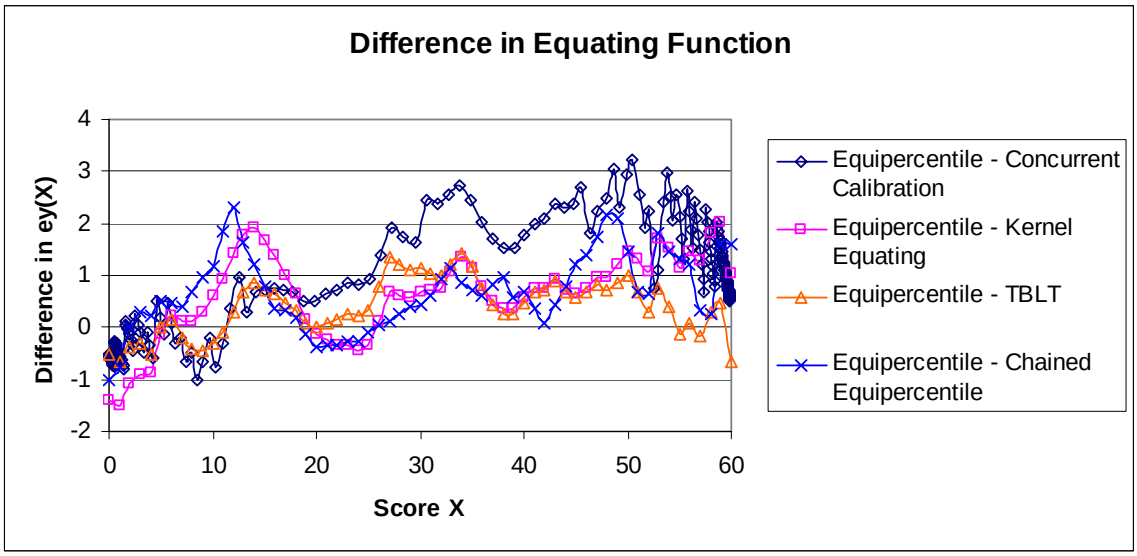


Figure F.40 60 Items per form, 50% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

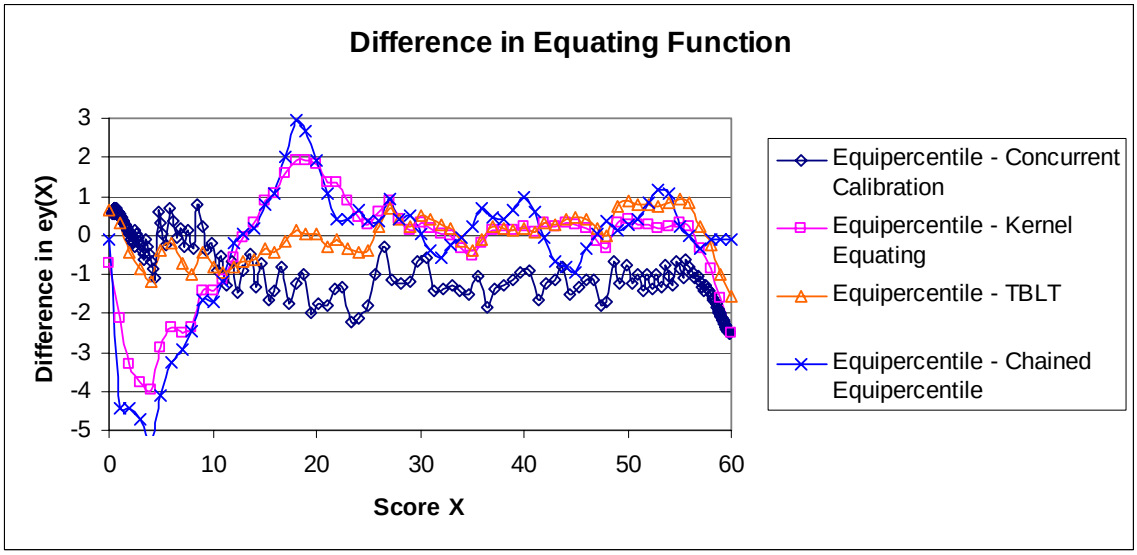


Figure F.41 60 Items per form, 50% Anchor Length, 10,000 Sample Size, No Theta Difference

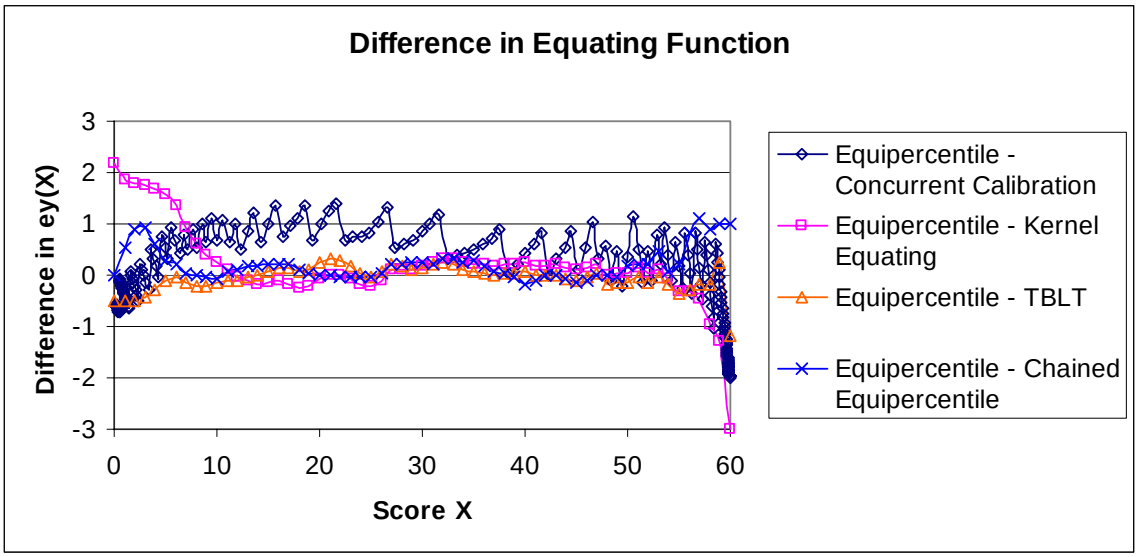


Figure F.42 60 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

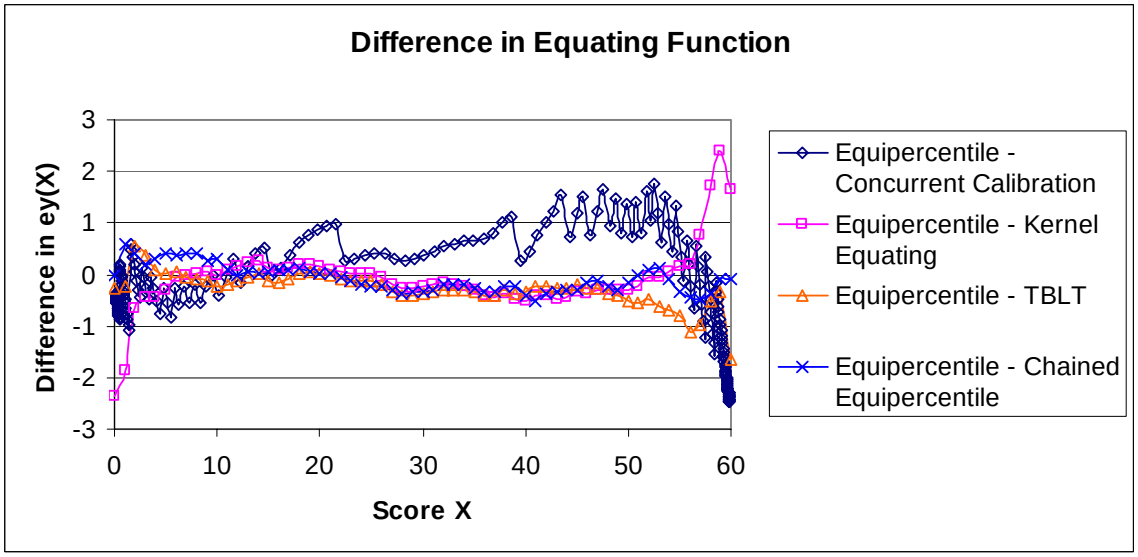


Figure F.43 60 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

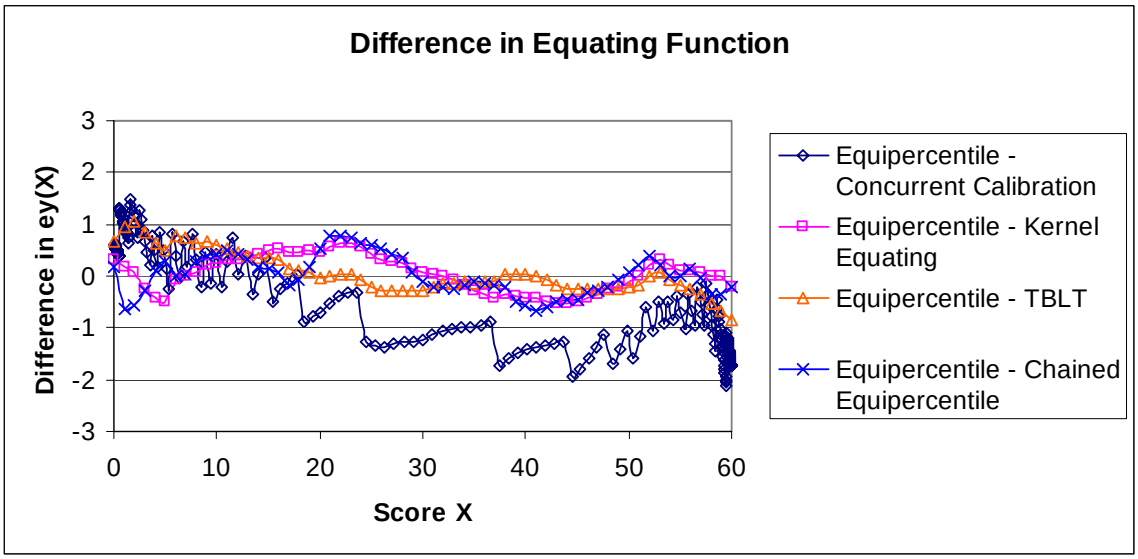


Figure F.44 60 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

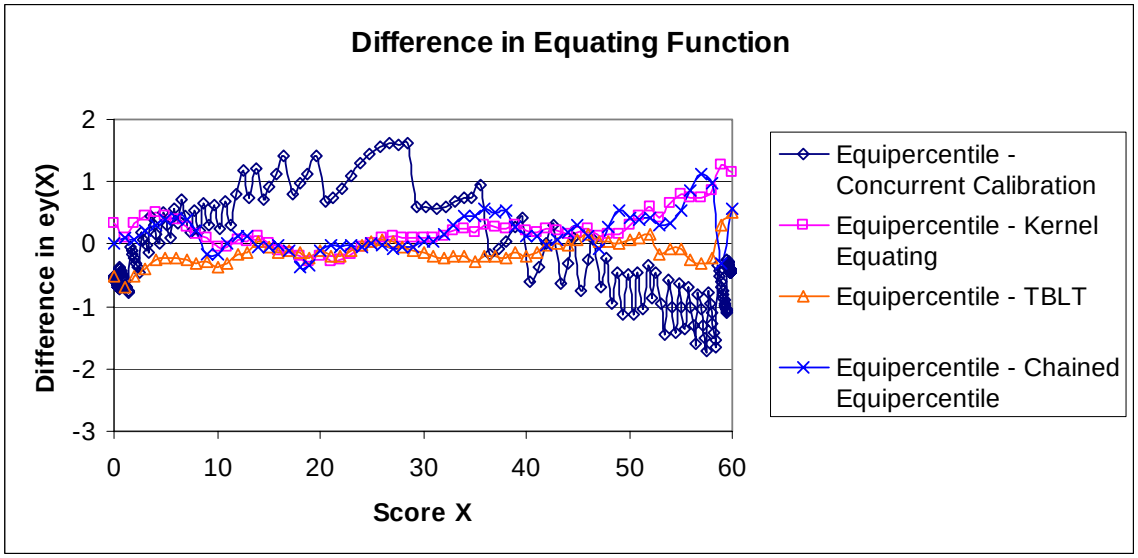


Figure F.45 60 Items per form, 50% Anchor Length, 100,000 Sample Size, No Theta Difference

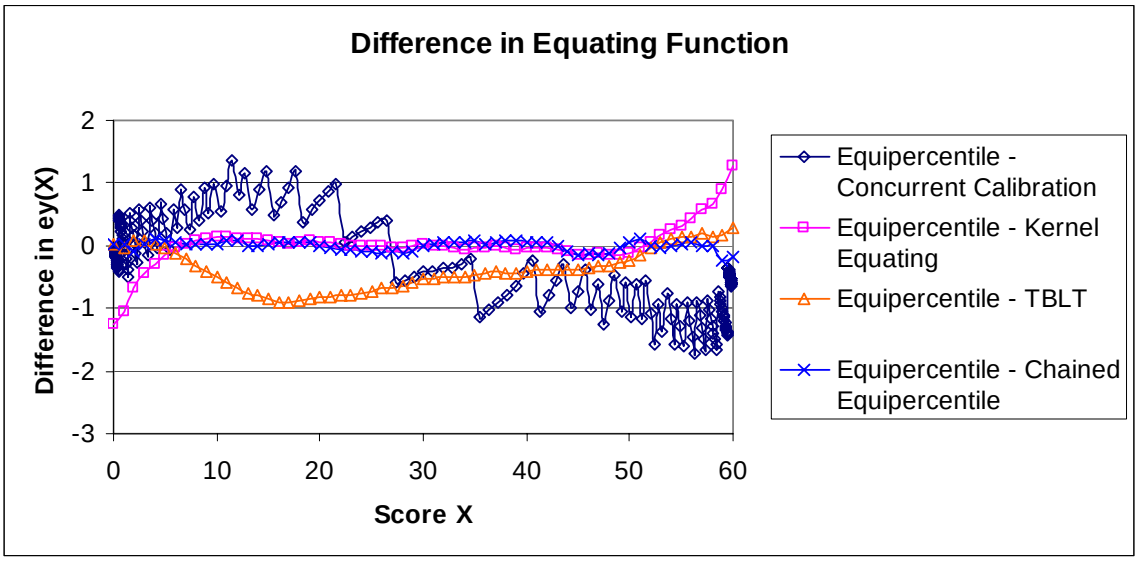


Figure F.46 60 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

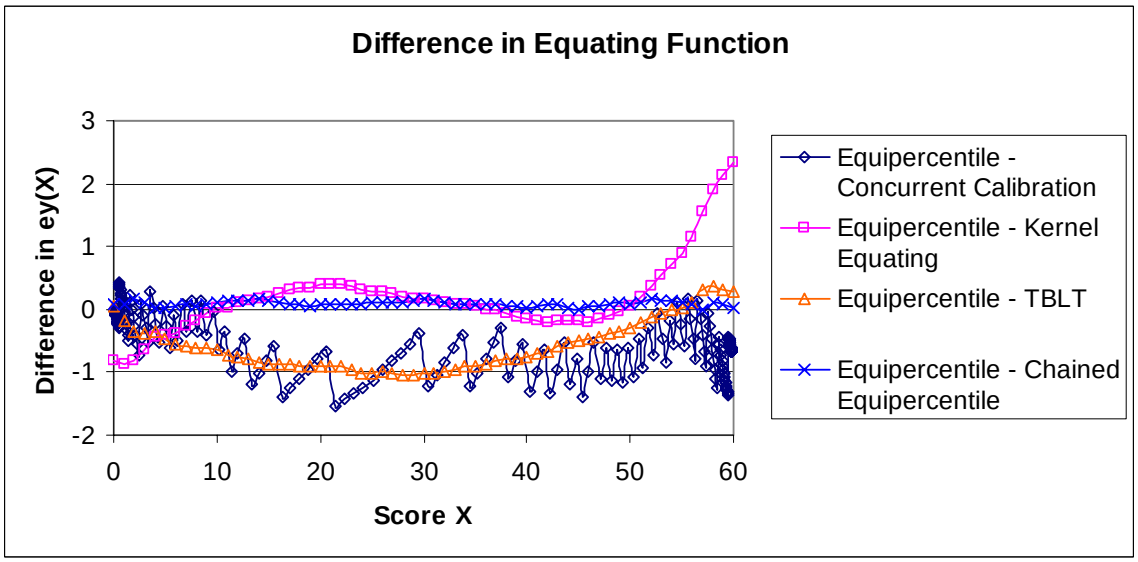


Figure F.47 60 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

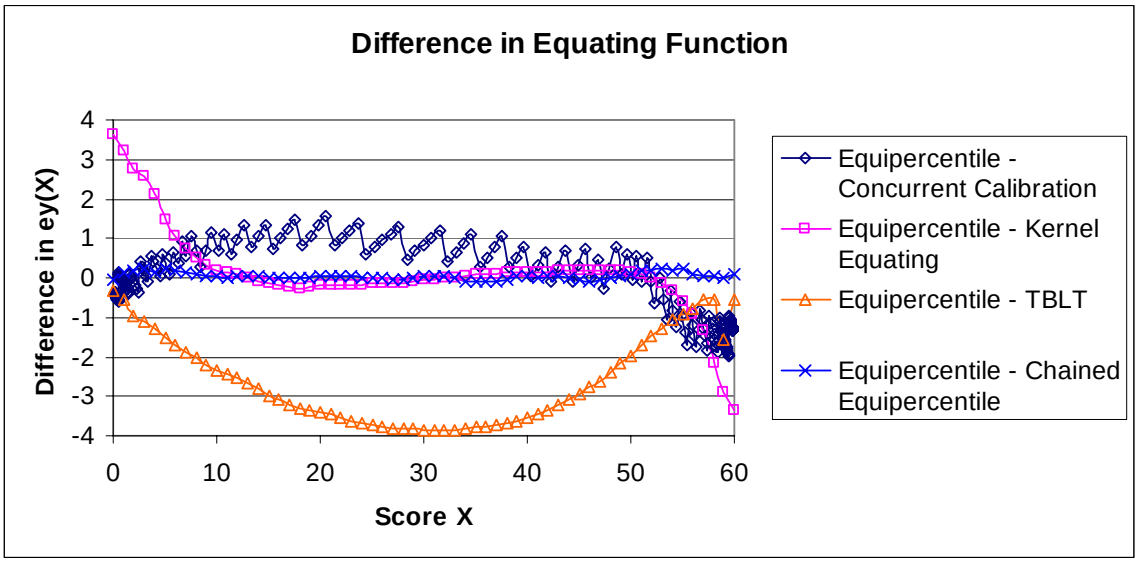


Figure F.48 60 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

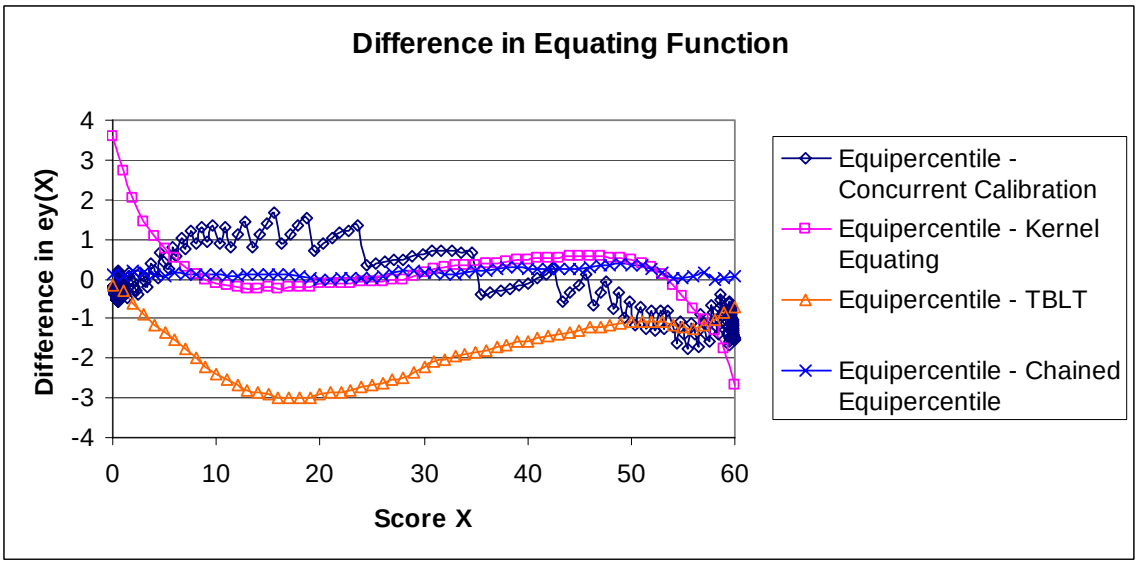


Figure F.49 60 Items per form, 35% Anchor Length, 1000 Sample Size, No Theta Difference

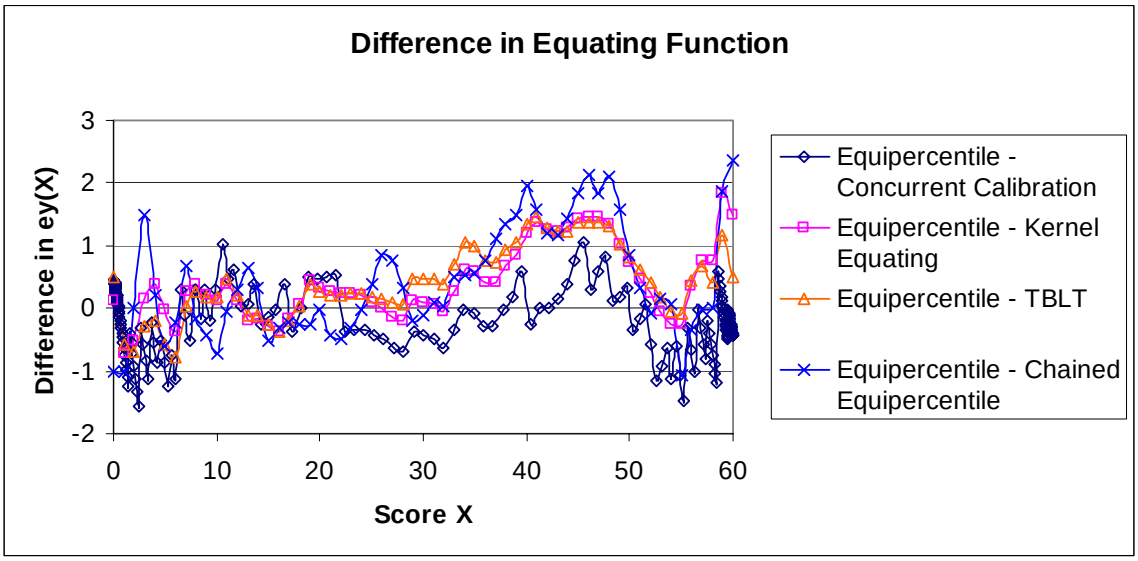


Figure F.50 60 Items per form, 35% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

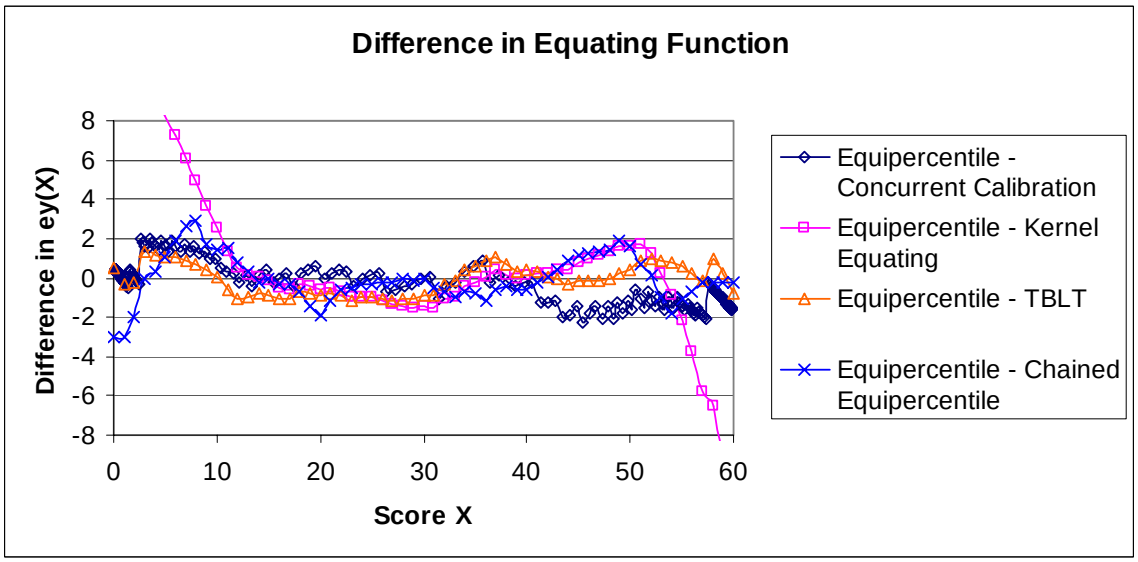


Figure F.51 60 Items per form, 35% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

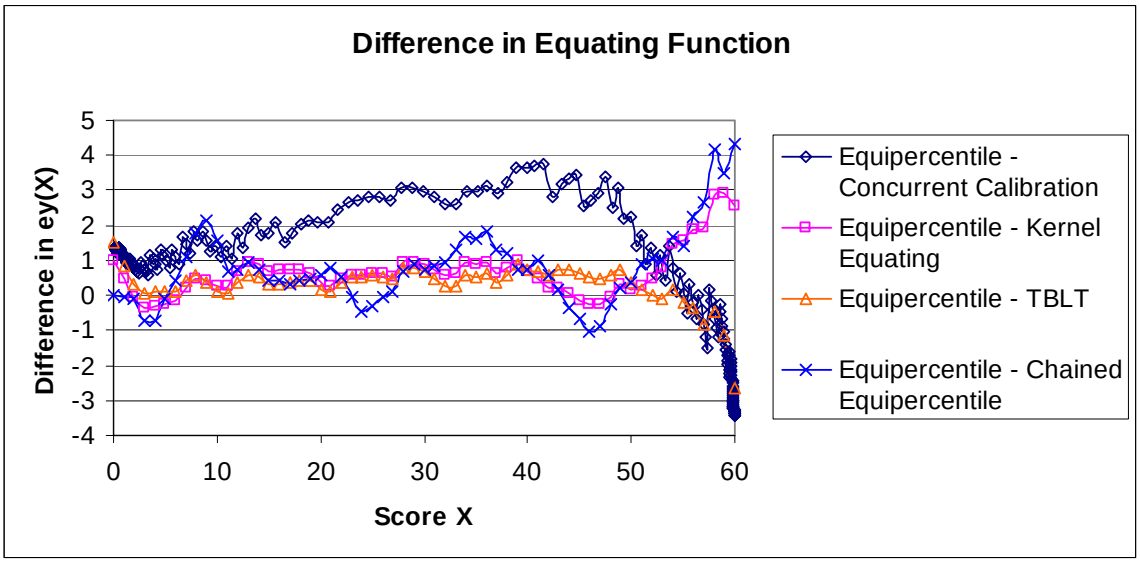


Figure F.52 60 Items per form, 35% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

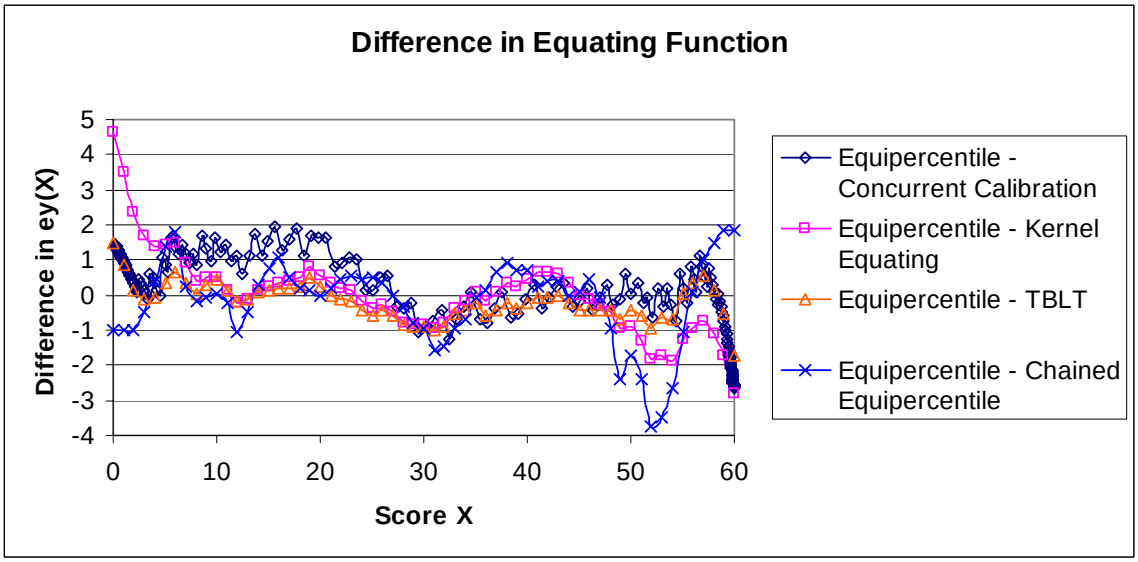


Figure F.53 60 Items per form, 35% Anchor Length, 10,000 Sample Size, No Theta Difference

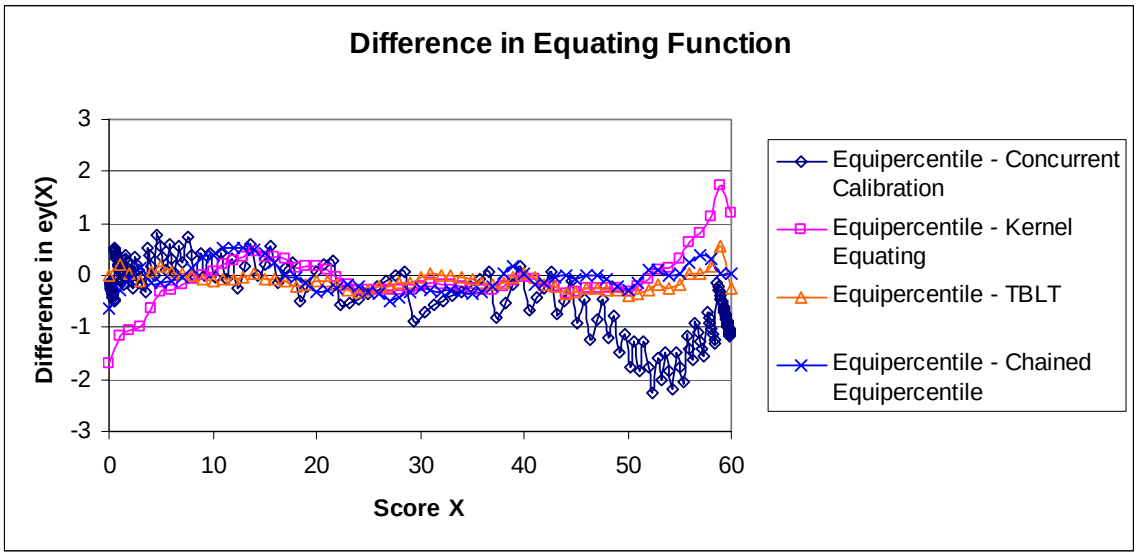


Figure F.54 60 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

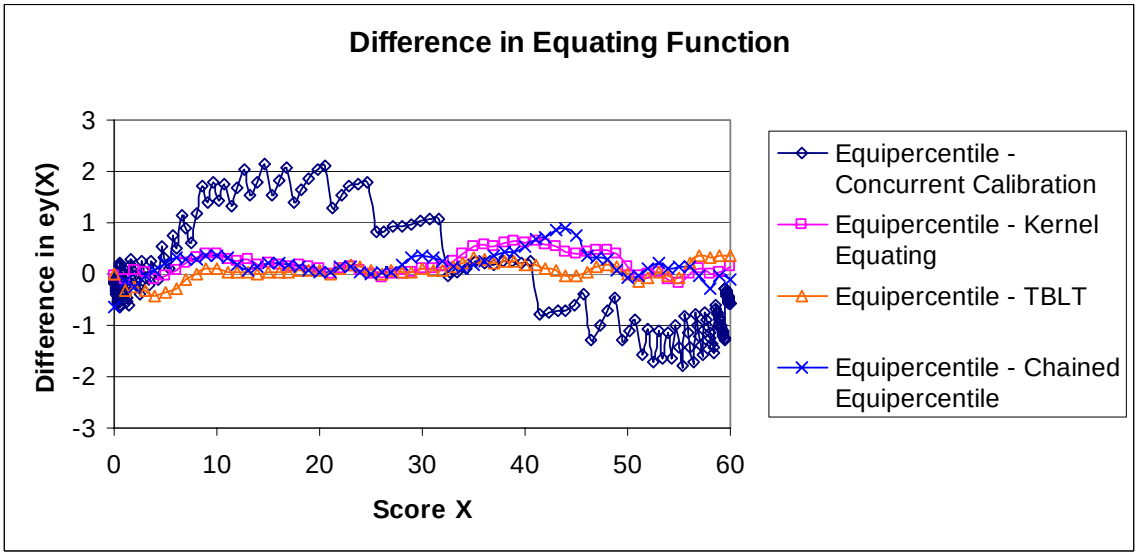


Figure F.55 60 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

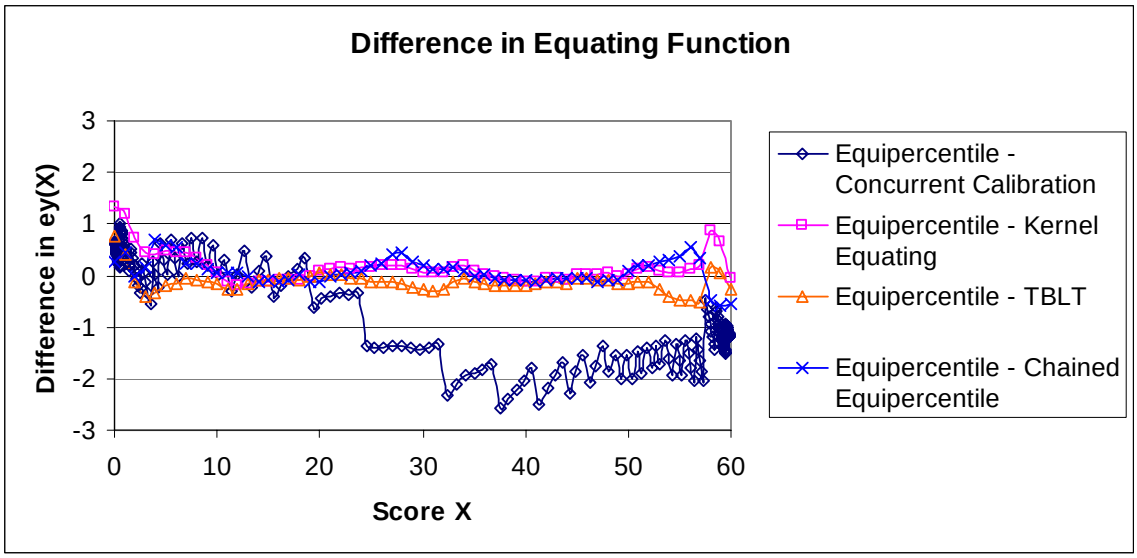


Figure F.56 60 Items per form, 35% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

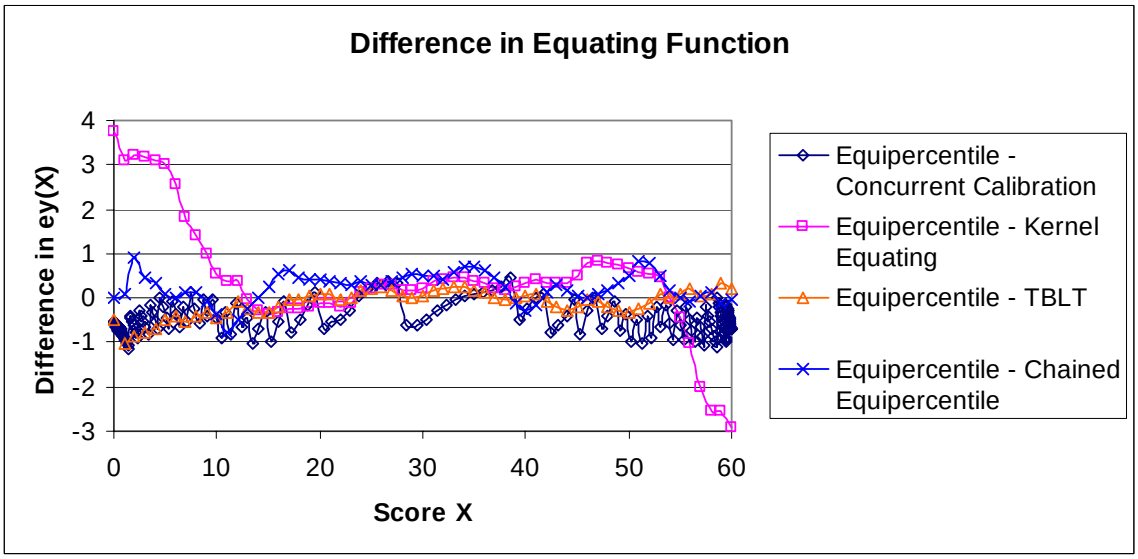


Figure F.57 60 Items per form, 35% Anchor Length, 100,000 Sample Size, No Theta Difference

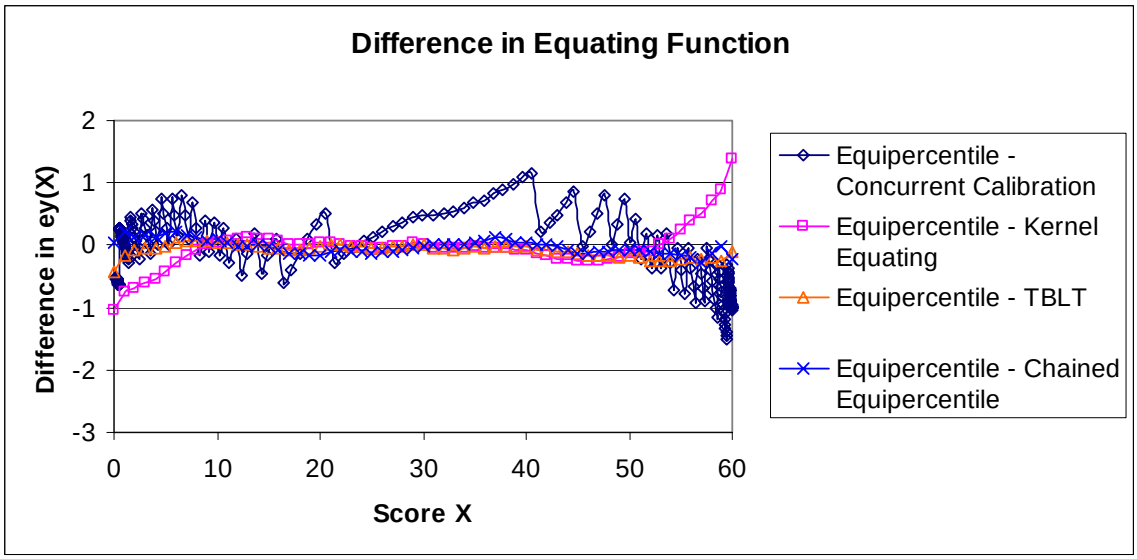


Figure F.58 60 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

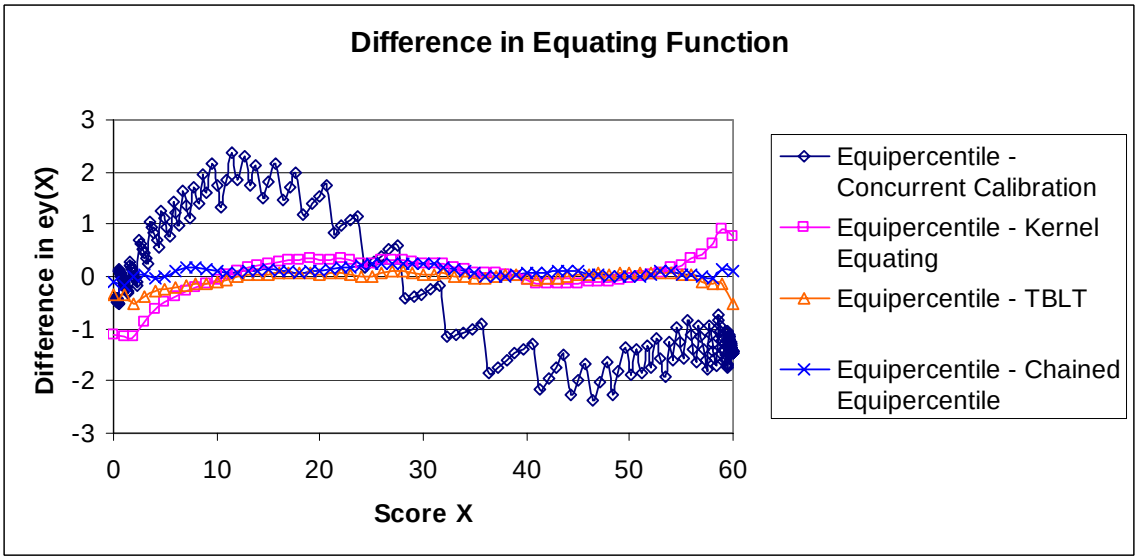


Figure F.59 60 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

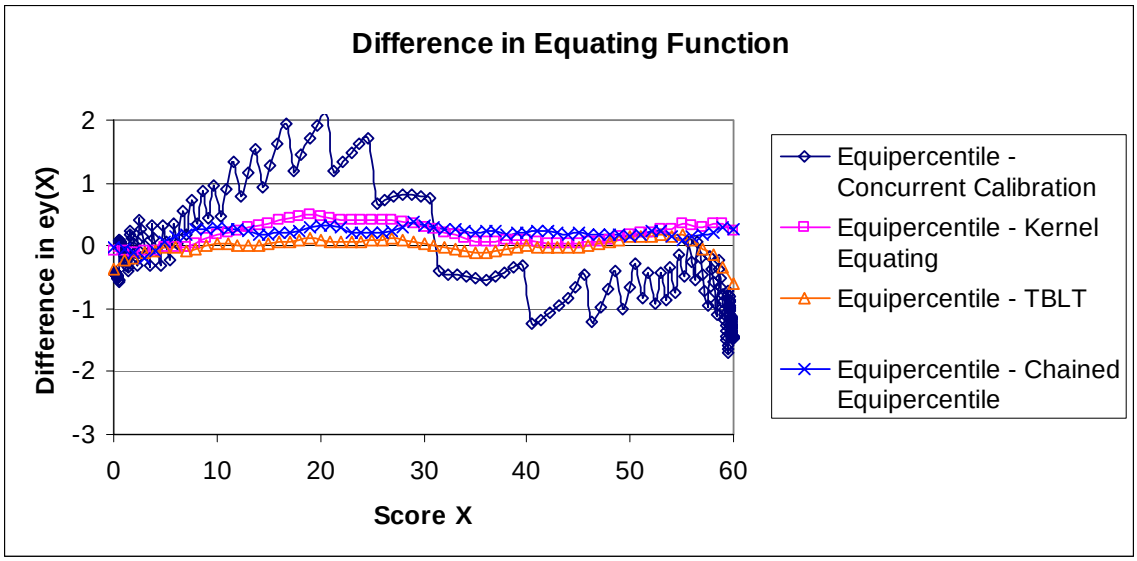


Figure F.60 60 Items per form, 35% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

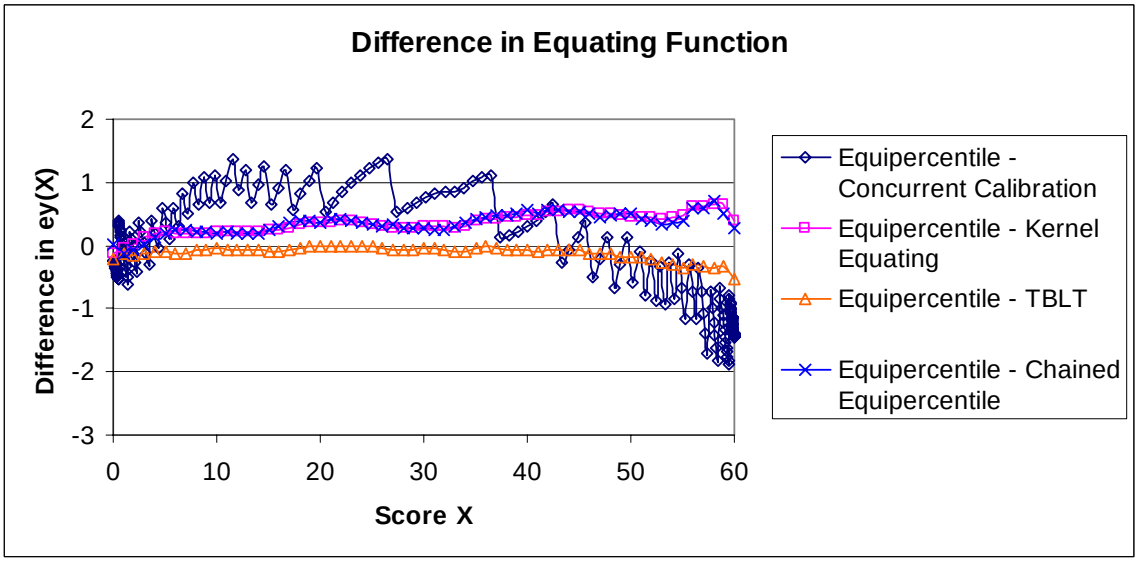


Figure F.61 60 Items per form, 20% Anchor Length, 1000 Sample Size, No Theta Difference

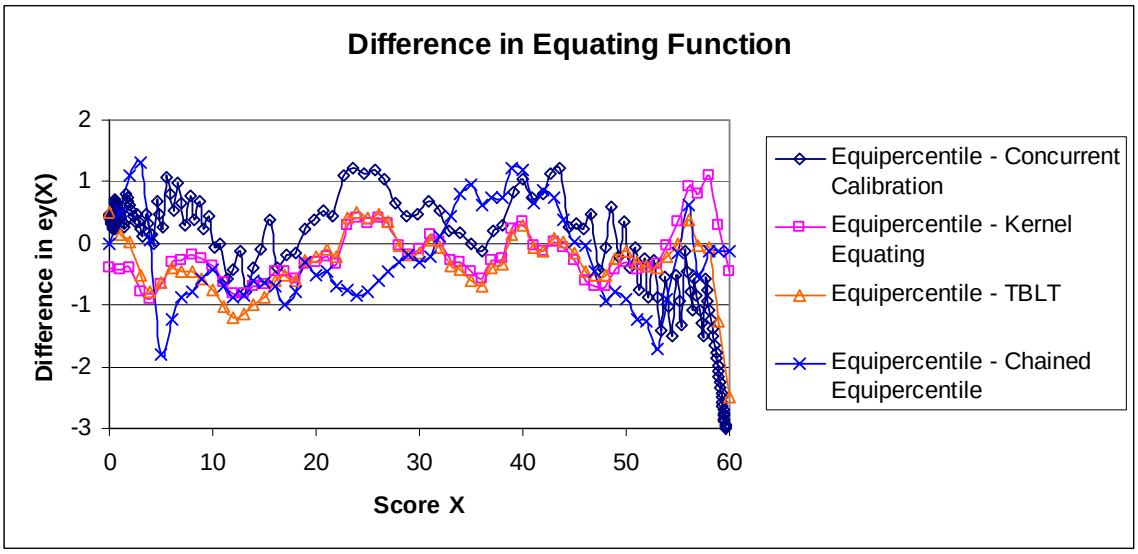


Figure F.62 60 Items per form, 20% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

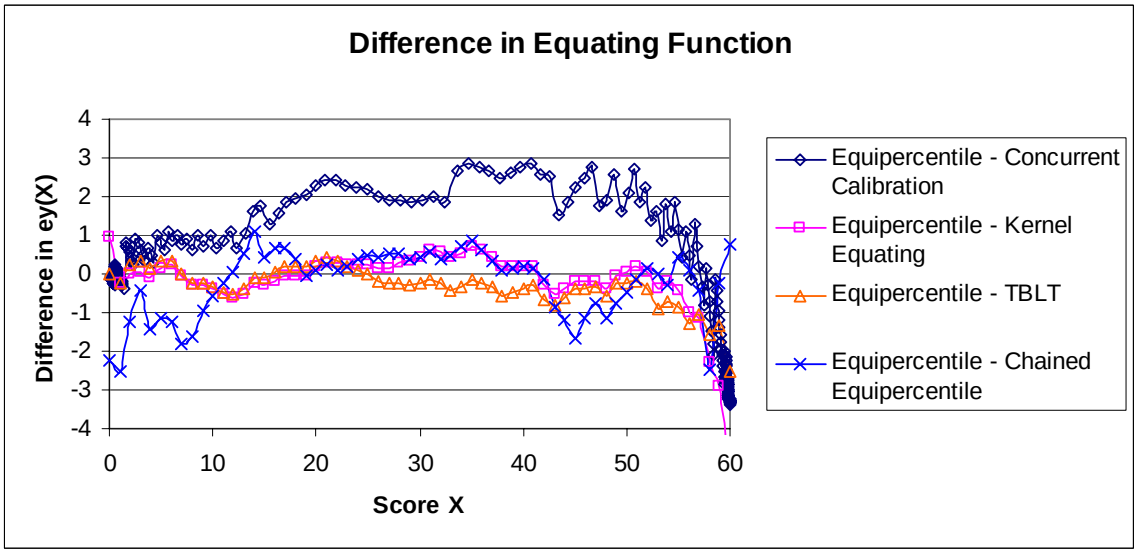


Figure F.63 60 Items per form, 20% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

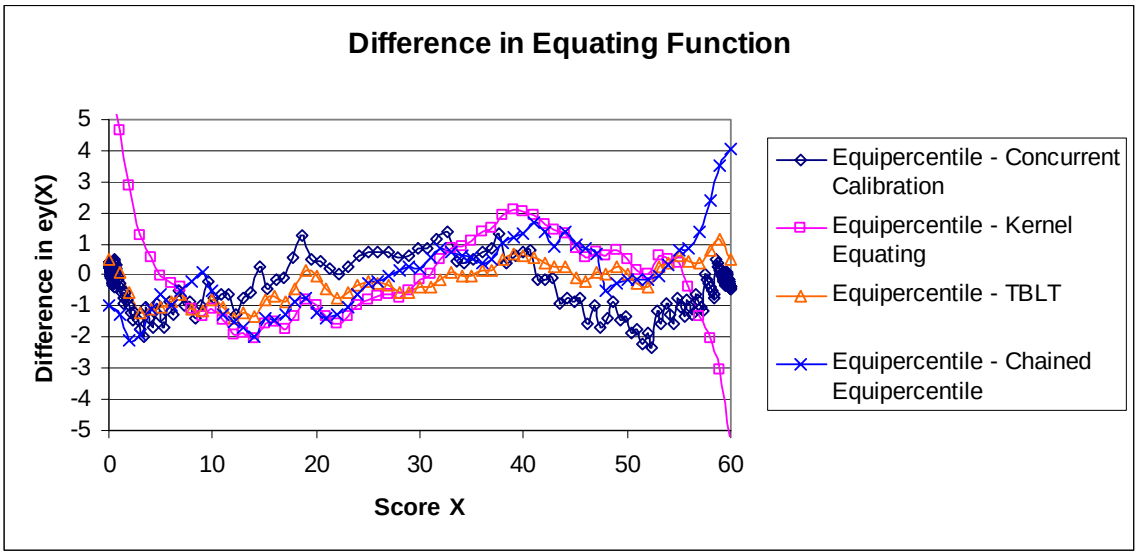


Figure F.64 60 Items per form, 20% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

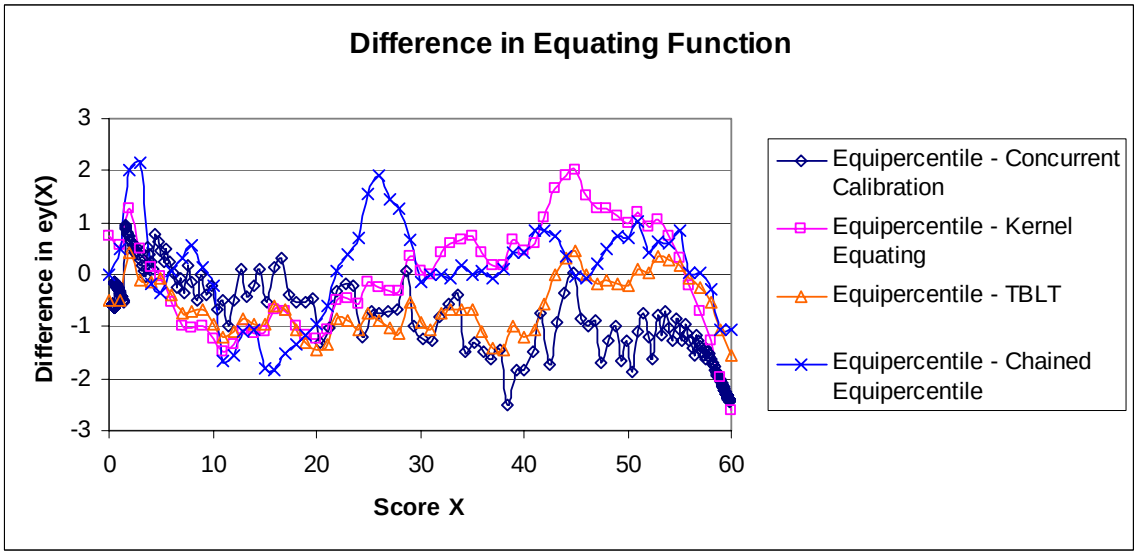


Figure F.65 60 Items per form, 20% Anchor Length, 10,000 Sample Size, No Theta Difference

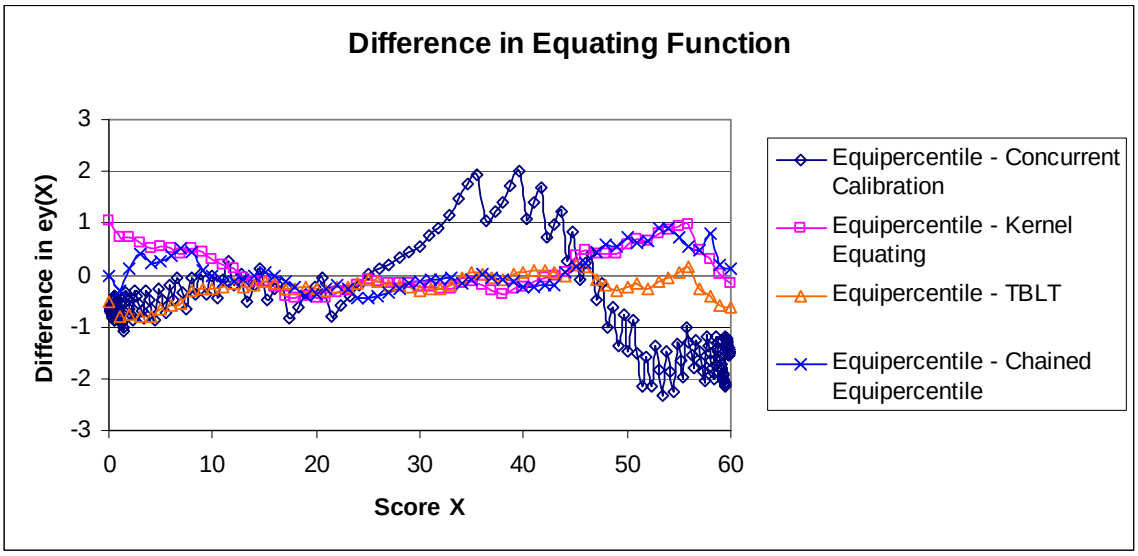


Figure F.66 60 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

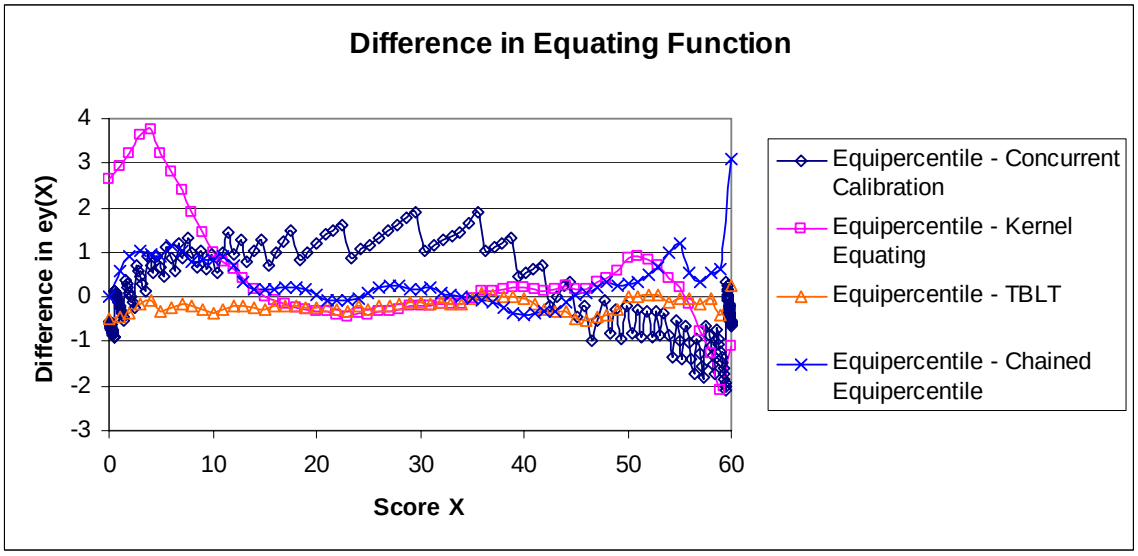


Figure F.67 60 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

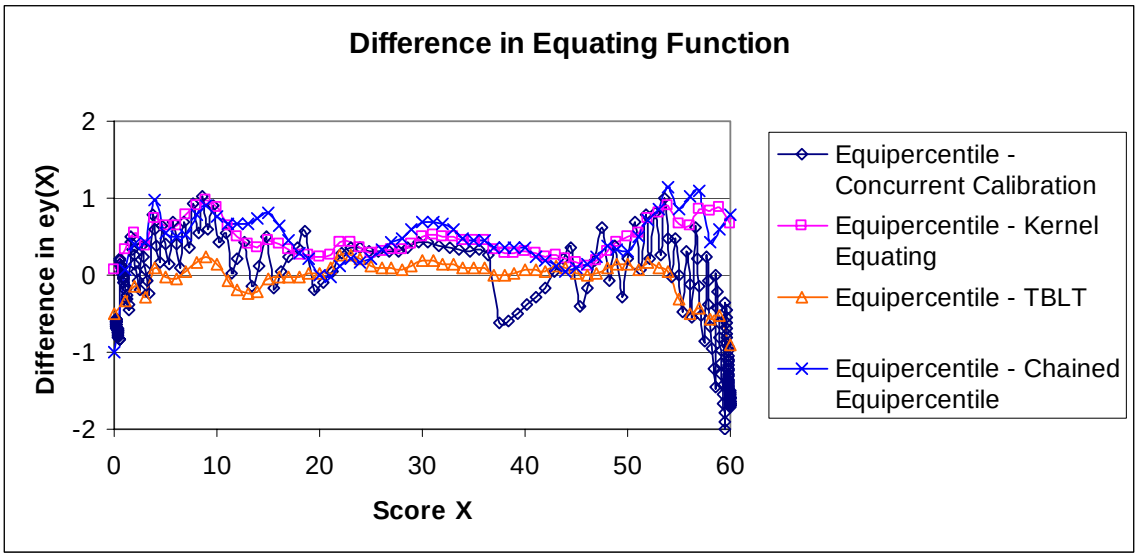


Figure F.68 60 Items per form, 20% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

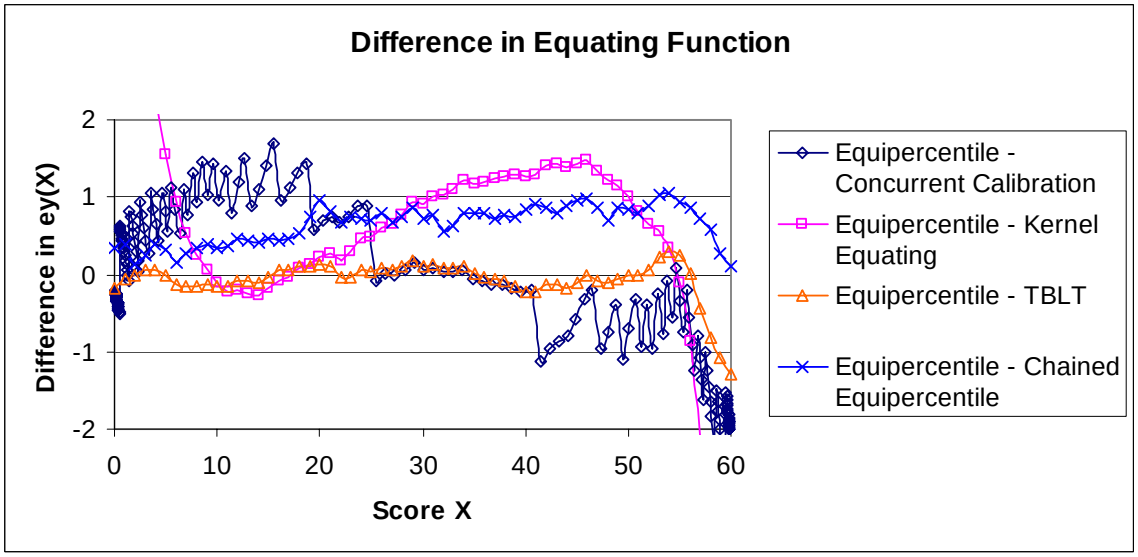


Figure F.69 60 Items per form, 20% Anchor Length, 100,000 Sample Size, No Theta Difference

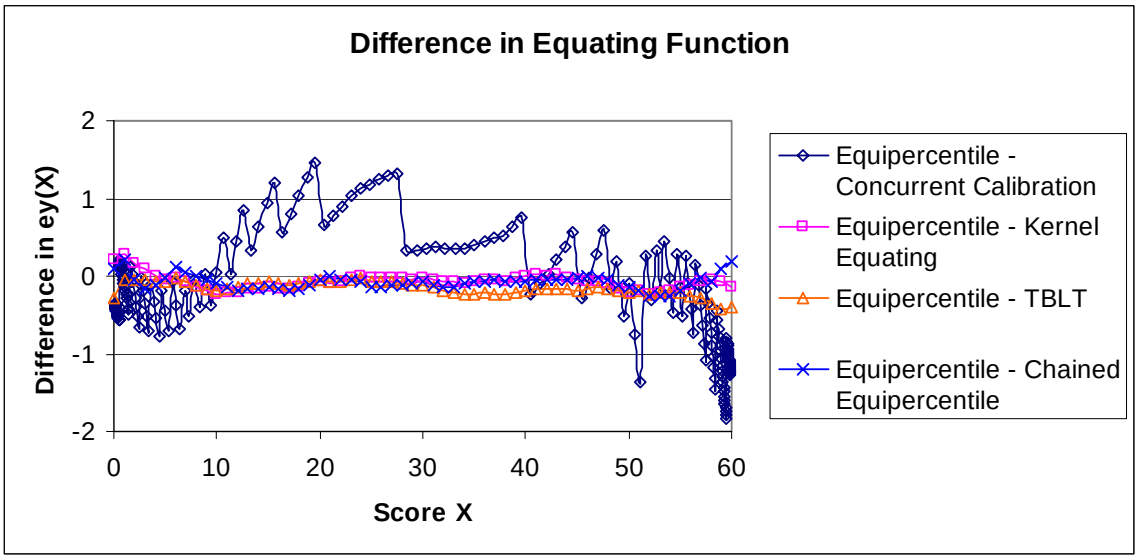


Figure F.70 60 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

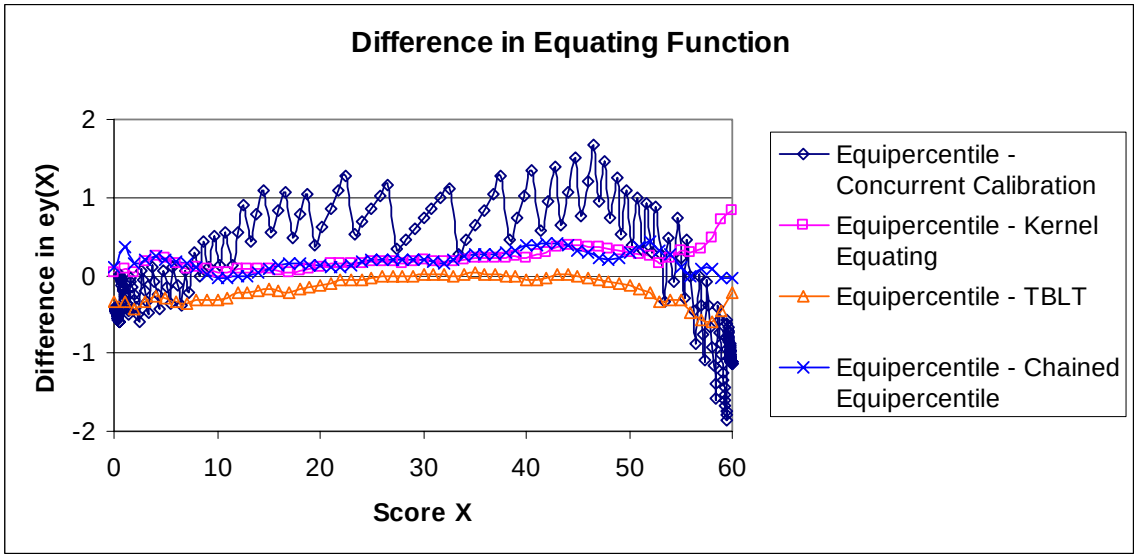


Figure F.71 60 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

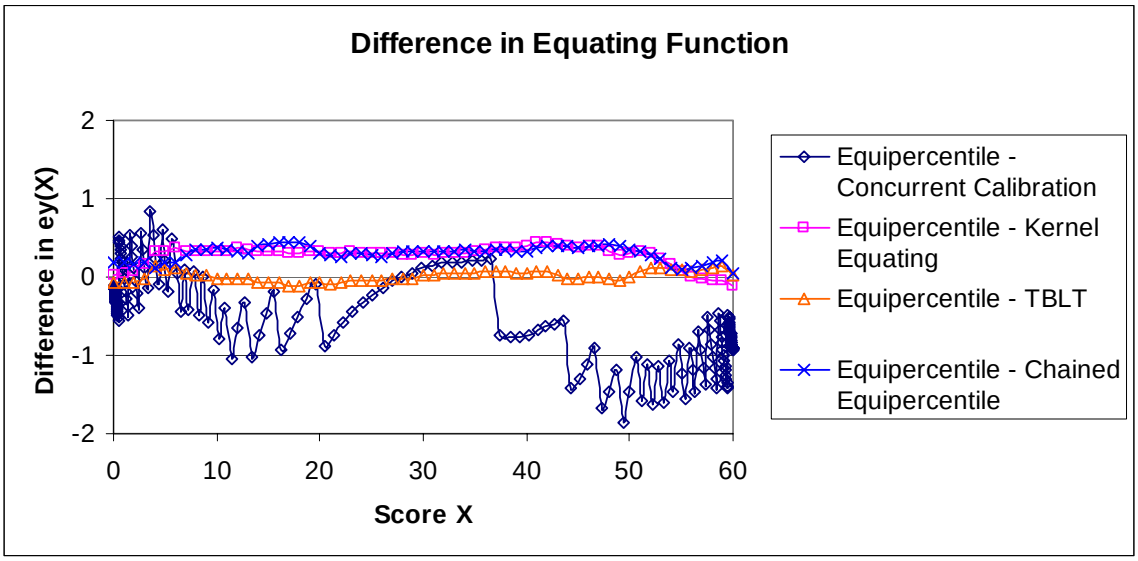


Figure F.72 60 Items per form, 20% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference

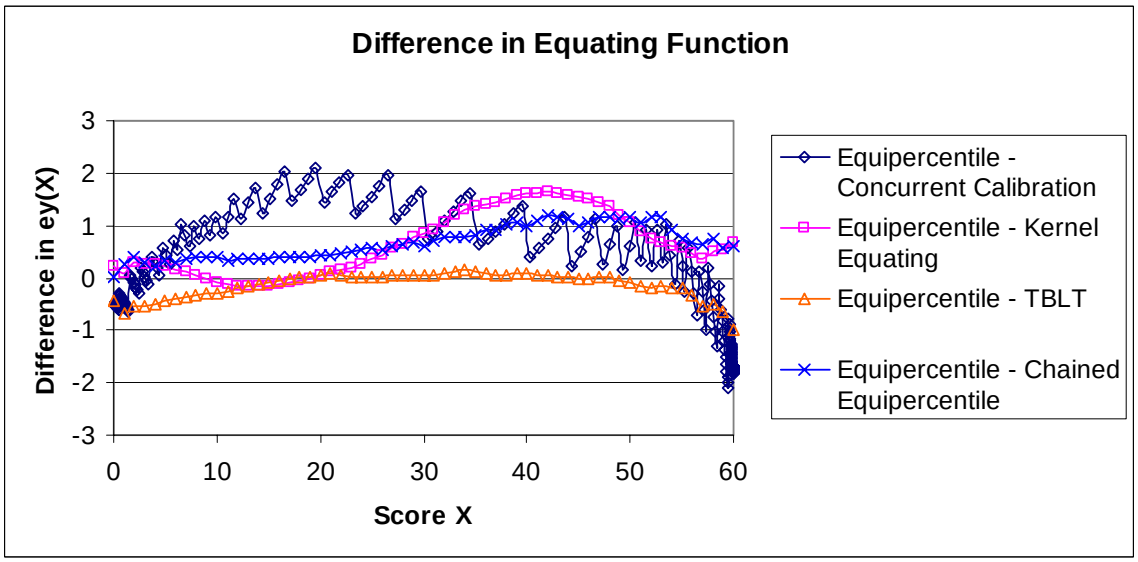


Figure F.73 20 Items per form, 50% Anchor Length, 1000 Sample Size, No Theta Difference

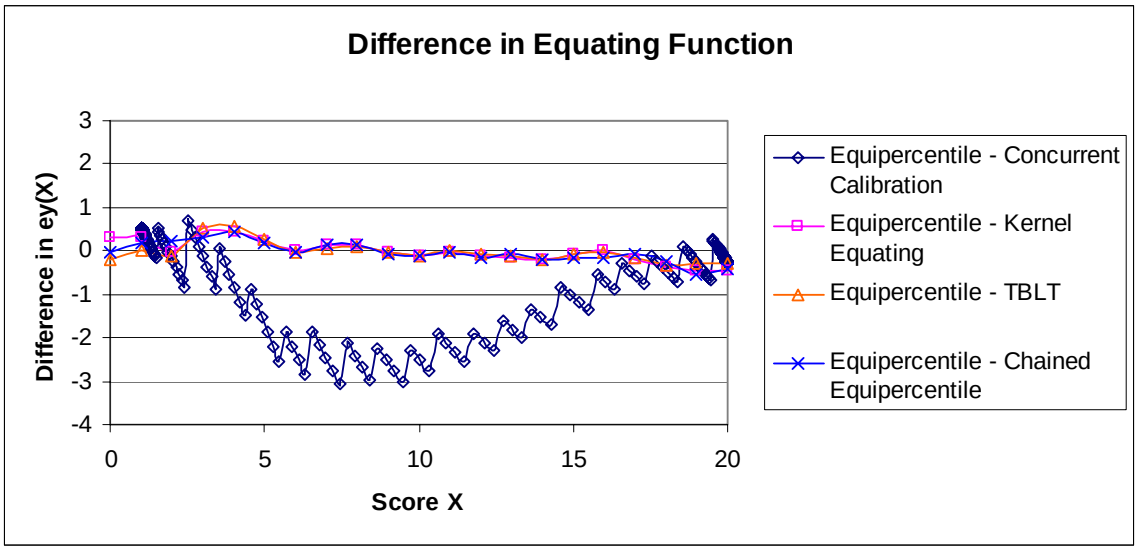


Figure F.74 20 Items per form, 50% Anchor Length, 1000 Sample Size, 0.1 Theta Difference

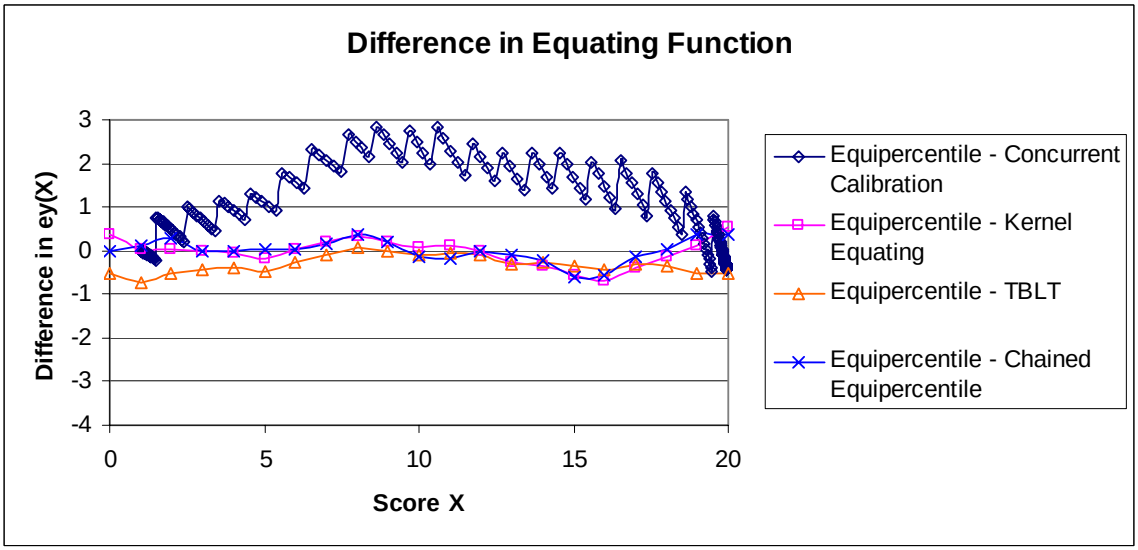


Figure F.75 20 Items per form, 50% Anchor Length, 1000 Sample Size, 0.2 Theta Difference

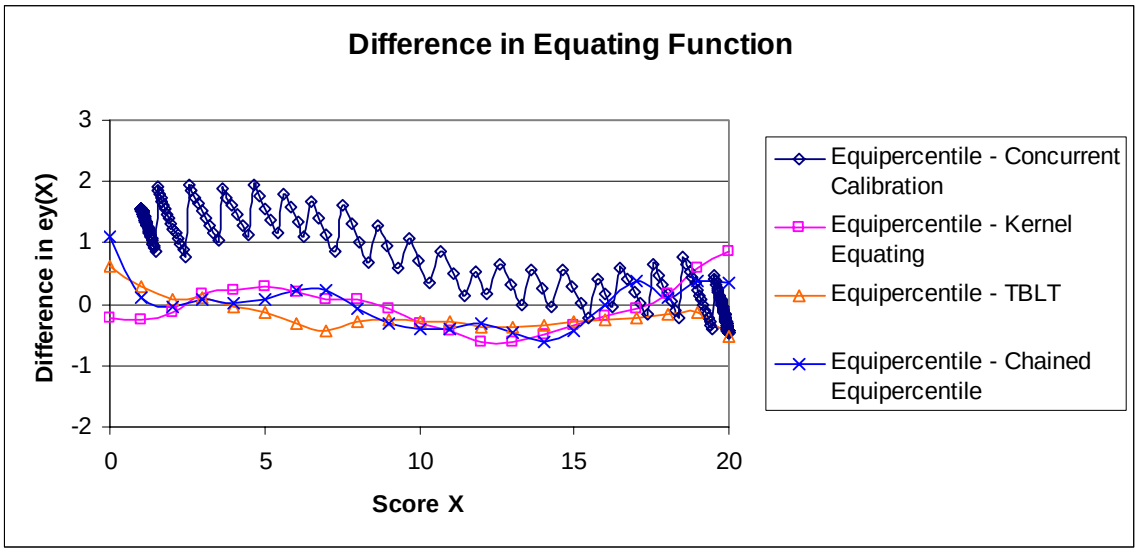


Figure F.76 20 Items per form, 50% Anchor Length, 1000 Sample Size, 0.4 Theta Difference

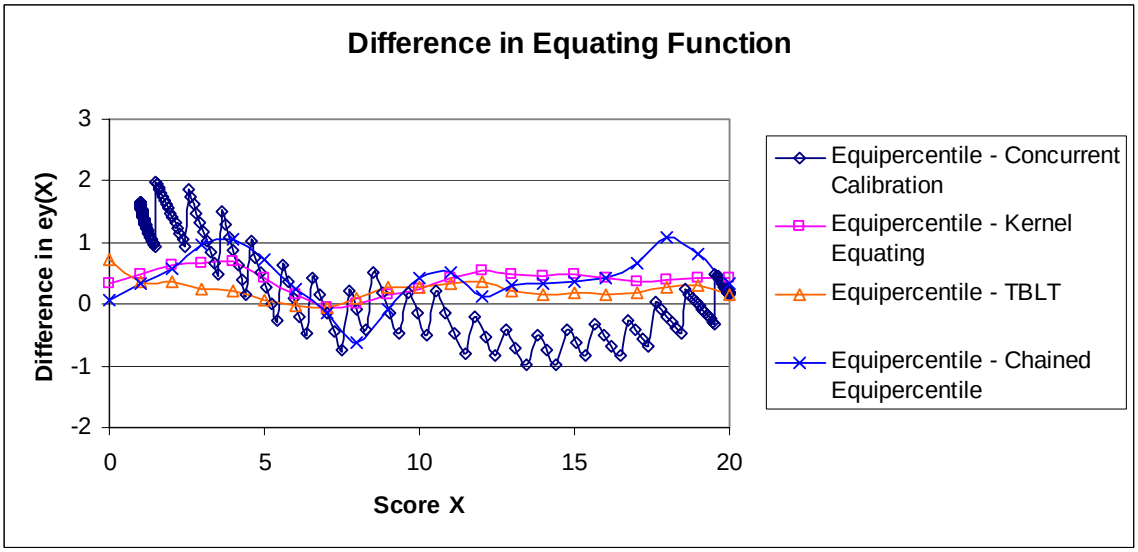


Figure F.77 20 Items per form, 50% Anchor Length, 10,000 Sample Size, No Theta Difference

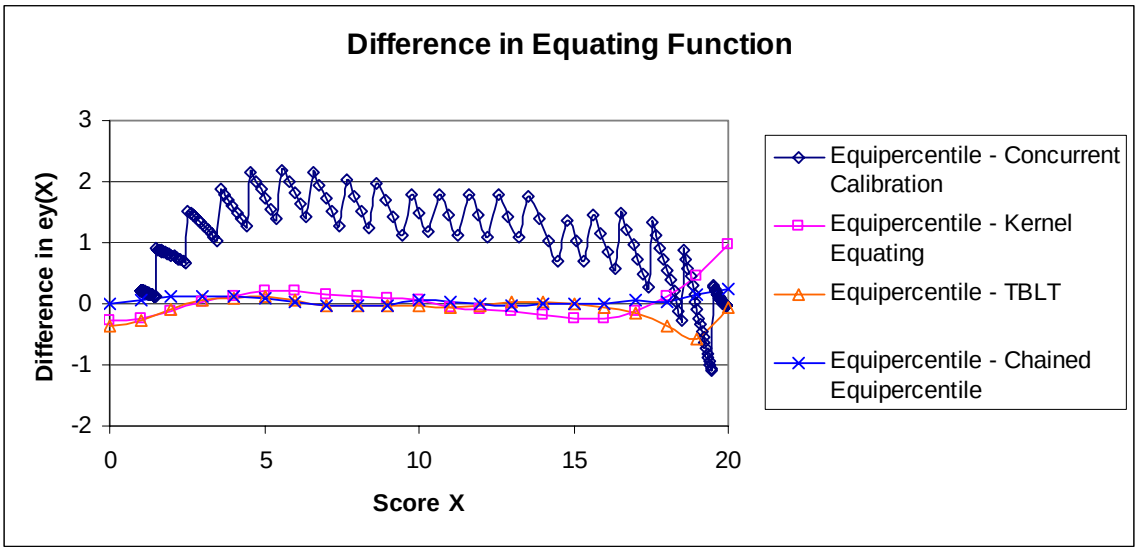


Figure F.78 20 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.1 Theta Difference

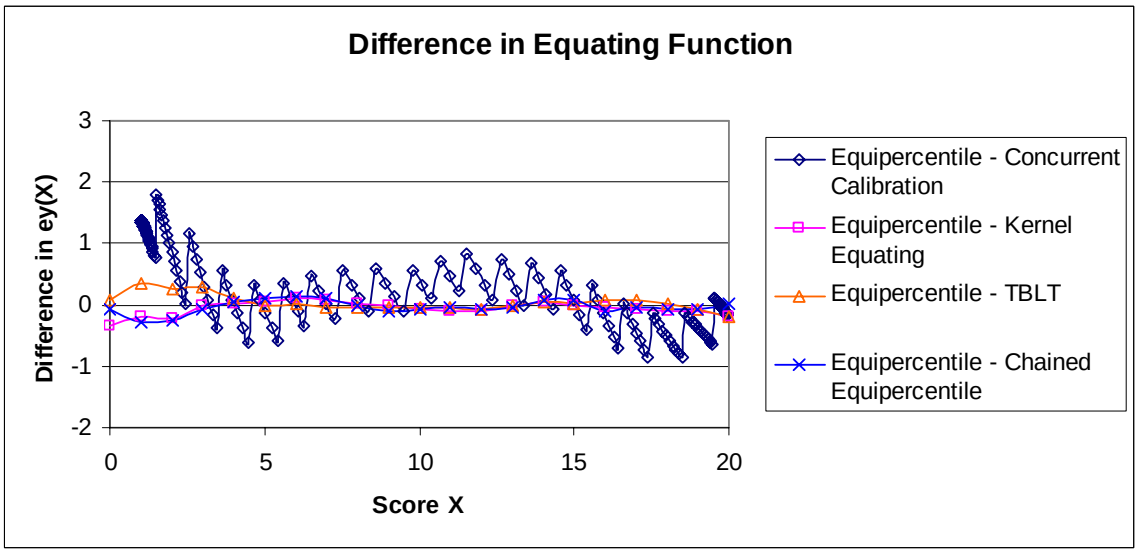


Figure F.79 20 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.2 Theta Difference

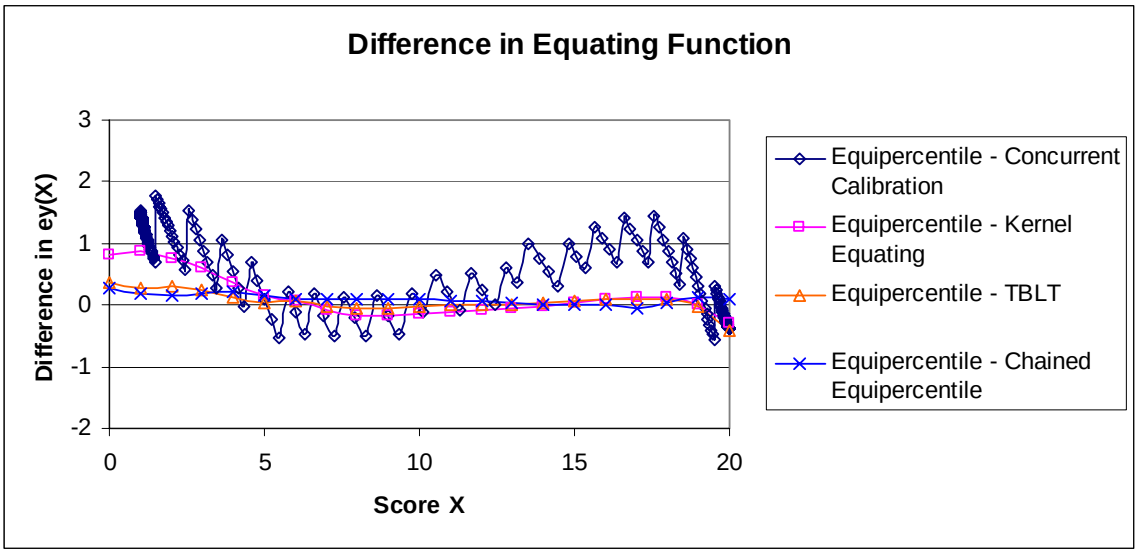


Figure F.80 20 Items per form, 50% Anchor Length, 10,000 Sample Size, 0.4 Theta Difference

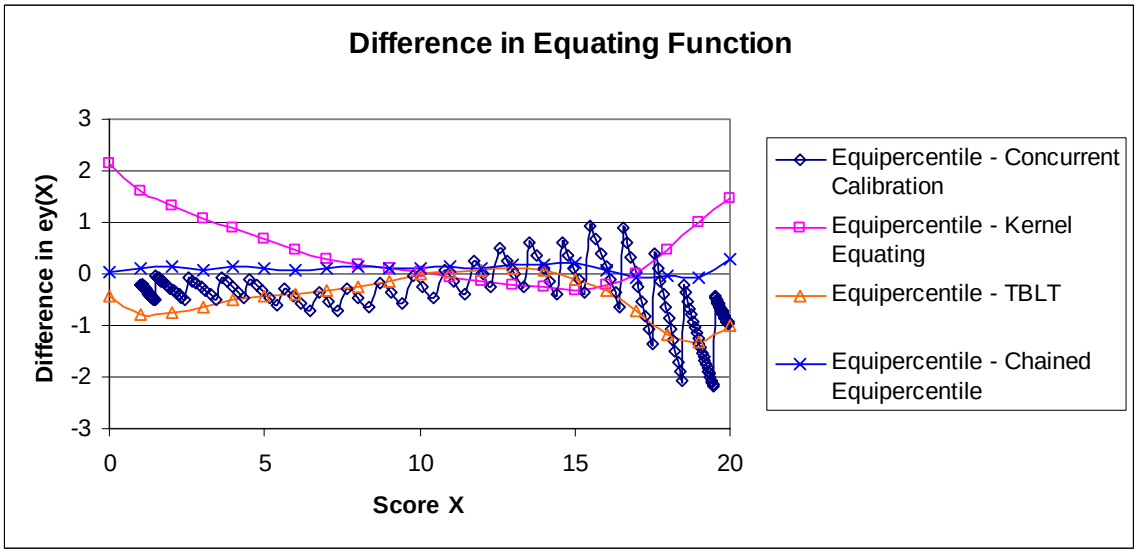


Figure F.81 20 Items per form, 50% Anchor Length, 100,000 Sample Size, No Theta Difference

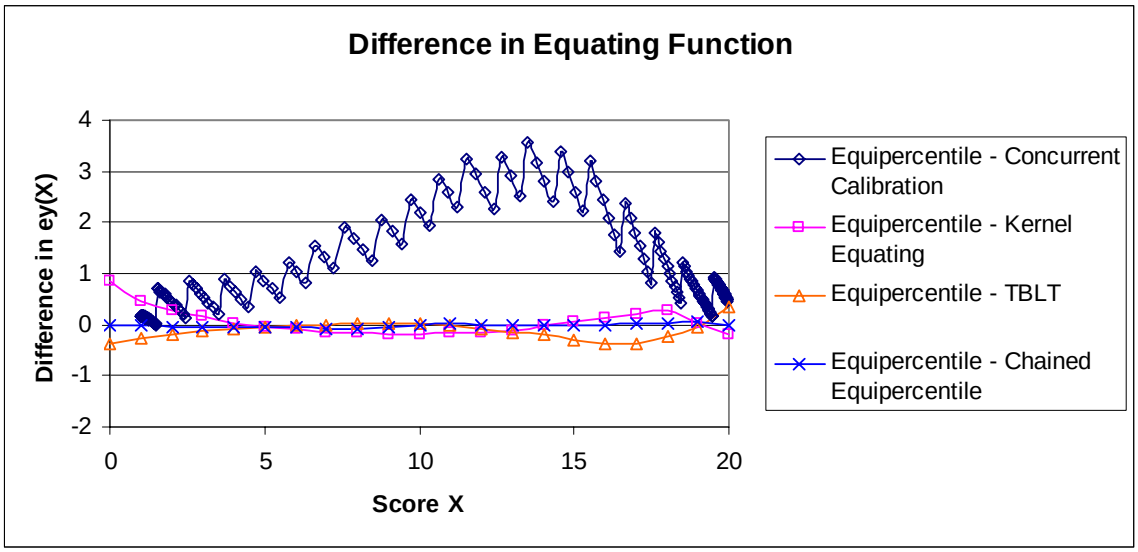


Figure F.82 20 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.1 Theta Difference

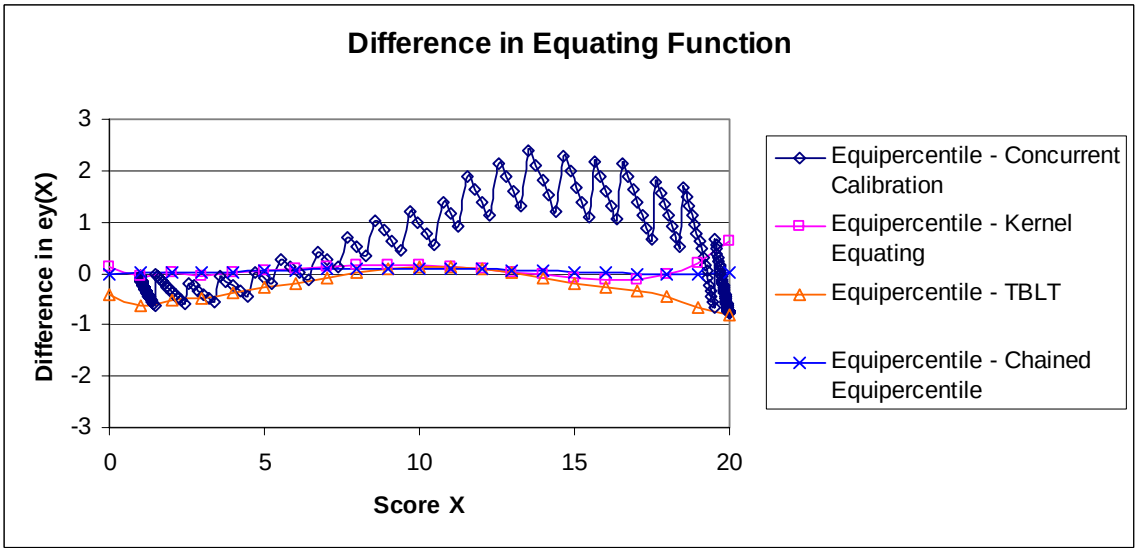


Figure F.83 20 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.2 Theta Difference

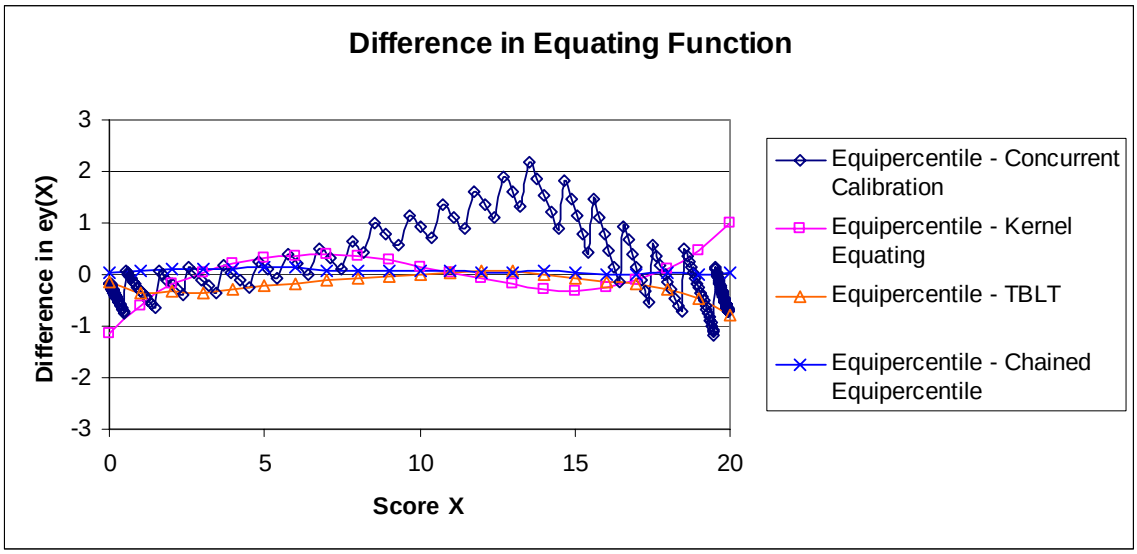


Figure F.84 20 Items per form, 50% Anchor Length, 100,000 Sample Size, 0.4 Theta Difference